UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Estimating caribou abundance in West Greenland using distance sampling methods

Iúri Josiane Frizado Correia

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Christine Cuyler
Professor Tiago André Marques

2020

Contents

List of Tables

List of Figures

Agradecimentos

Em primeiro lugar quero agradecer aos meus dois orientadores, Tiago Marques e Christine Cuyler. Ao Tiago pela calma aquando da tempestade e pela maneira relaxada de trabalhar, extremamente compatível com a minha. À Christine pela grande disponibilidade e conhecimento fornecidos ao longo do trabalho, bem como a empatia ao longo do mesmo. A ambos, a grande oportunidade de ter participado neste projeto que me fez crescer em muitas vertentes.

Também quero agradecer aos meus amigos chegados, estando em Portugal ou não, e aos meus familiares que estiveram (e aos que continuam) presentes durante todo este percurso, nos bons e maus momentos, perto ou à distância de uma chamada ou mensagem. Um agradecimento também aos meus colegas e corpo docente do curso de Mestrado em Bioestatística, do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências e do Centro de Estatística e Aplicações da Universidade de Lisboa pelo ambiente em que estive incluído e pela enorme aprendizagem nestes últimos anos.

Um miminho especial para a minha colega Beatriz Afonso, pela ajuda crucial e pelas horas e paciência disponibilizadas, à minha tia Irene pelas horas ao telefone e pela companhia e, claro, às Donas e às minhas amigas Eva Alves, Mónica Louro, Mónica Martins, Mariana Martins e Patrícia Felício por simplesmente serem quem são.

Por fim, e mais importante, ao meu quarteto fantástico e mais que tudo, mamele Neuza e papele Sérgio pelo apoio e amor incondicional todos os dias, aquele carinho que só eles podem e sabem dar, à minha irmã Quéssia por nunca se cansar de mim e saber-me ler de qualquer forma, e à minha irmã Teresinha por fazer questão de me distrair enquanto descansava na sala, nos fins de semana de visita a casa. Palavras não chegam para descrever esta família. É a eles que dedico este trabalho de final de curso, porque sempre foram e sempre serão a minha motivação. Família é família! Família é família mesmo…

— Iúri J. F. Correia

Resumo

O caribu, *Rangifer tarandus groenlandicus*, é uma das treze subespécies de rena que habita as regiões circumpolares do hemisfério Norte, sendo nativa da Gronelândia. A maioria dos estudos de investigação feitos no passado foram, regra geral, de curta duração, limitados e frequentemente mal desenhados ou extremamente enviesados. Ao ser caçado há pelo menos 4000 anos, o caribu possui elevada importância para as populações locais. Como tal, métodos com maior eficiência e precisão precisam de ser considerados para que a informação e monitorização obtidas para esta subespécie sejam mais sólidas e robustas.

Desde o início do milénio, o Instituto para os Recursos Naturais da Gronelândia (Greenland Institute for Natural Resources - GINR) tem vindo a melhorar as metodologias utilizadas no controlo e monitorização das espécies presentes na Gronelândia, sendo o caribu uma das espécies de foco. No Oeste da Gronelândia, local onde existem populações nativas de caribus, não existem predadores naturais, o que torna as próprias características populacionais inerentes desta subespécie, características do habitat e a atividade antropogénica nos principais fatores responsáveis pelo controlo da população.

A partir do ano de 2000, o GINR tem feito estudos aéreos de monitorização periódica do *R. t. groenlandicus*. O intuito destes estudos é o de estimar a sua densidade e abundância nas regiões de interesse, obtendo assim um panorama geral da sua abundância, distribuição e relação com algumas covariáveis recolhidas. Posteriormente, os seus resultados são usados para auxiliar a gestão de outras atividades como a caça, construção, entre outras. Recorriam-se a vários métodos baseados em estudos mais antigos mas, mais recentemente, os métodos de amostragem por distâncias têm vindo a ganhar destaque na investigação ecológica, tanto pela sua aplicabilidade em múltiplas situações como pela precisão das estimativas fornecidas.

Primeiramente foi feita uma análise exploratória dos dados. Esta revelou que foram detetados cerca de 5 000 caribus em grupos maioritariamente até 4 indivíduos e localizados na sub-área de Sisimiut. Também mostrou ser concordante com os resultados obtidos em estudos anteriores e sugeriu elevada precisão inerente à amostra observada. Posteriormente procedeu-se à aplicação de métodos de amostragem por distâncias e ajustamento de um modelo espacial.

Os métodos de amostragem por distâncias são usados para a estimação da densidade e abundância de objetos de interesse com base na distância dos mesmos ao observador. Esse objeto de interesse pode ser um ser vivo como uma espécie de animal ou de planta, ou pistas produzidas pelos mesmos (maioritariamente em animais) como ninhos, carcaças, fezes ou até

mesmo sons. Aplicou-se este tipo de metodologia onde, essencialmente, os observadores, quer no campo quer no ar ou em alto-mar, se deslocam ao longo de um transeto ou permanecem estáticos num determinado ponto por um intervalo de tempo fixo, observando a área em redor. Sempre que um objeto de interesse, no caso deste estudo, caribus individuais ou em grupo, é detetado, as distâncias perpendiculares dos mesmos à linha central do transeto são medidas e registadas. Outras covariáveis consideradas importantes na deteção dos objetos de interesse também podem ser recolhidas no momento da amostragem e analisadas em conjunto com as distâncias osbervadas. Estas distâncias permitem depois estimar uma função de deteção, a partir da qual se pode obter uma estimativa da probabilidade de deteção dos objetos de interesse. Esta estimativa representa a proporção de objetos de interesse detetados durante a amostragem e a sua relação com a distância ao observador é expressa pela função de deteção. Posteriormente, pode-se estimar a densidade e abundância totais ou estratificadas em caso de existência de sub-áreas dentro da área em estudo, o que acontece neste projeto.

Após a seleção do modelo de amostragem por distâncias mais apropriado à amostra recolhida, procedeu-se ao ajustamento de um modelo aditivo generalizado (Generalized Additive Model - GAM) para determinar o número de caribus estimado em cada quadrícula de $1.5km \times 1.5km$, considerando a probabilidade de deteção estimada anteriormente e a influência de outras covariáveis ambientais georreferenciadas também recolhidas e de interesse *a priori*. As covariáveis ambientais consideradas neste projeto foram o aspeto geográfico, a elevação e o declive.

Recorrendo aos resultados do GAM obtido, um modelo de densidade espacial (Density Surface Model - DSM) dos objetos de interesse foi determinado através da extrapolação para a região de estudo. Este modelo representa a distribuição dos caribus ao longo da região de estudo como função das várias covariáveis consideradas sob a forma de um mapa de gradiente de cor. Para complementar o DSM, também foi criado um segundo mapa de gradiente de cor onde cada quadrícula representa a incerteza associada a cada estimativa obtida no DSM.

Relativamente à distribuição espacial estimada pelo DSM para a espécie de estudo, este prevê que o caribu se encontra maioritariamente na sub-área de Sisimiut e minoritariamente na sub-área de Angujaartorfiup Nunaa, como previsto pela especialista. De acordo com o modelo selecionado, em termos de densidade, os caribus aparentam preferir superfícies direcionadas para Sul em detrimento das direcionadas para Norte. Além disto, a espécie de estudo parece apresentar uma relação negativa com o aumento da elevação e do declive. Estes resultados são concordantes com os obtidos em estudos anteriores realizados na mesma altura do ano.

Para o presente projeto, focado na região Norte do Oeste da Gronelândia, a análise de amostragem por distâncias estimou uma probabilidade de deteção média de 0.54 e uma estimativa de abundância com base no delineamento amostral de cerca de 60 469 (IC 95%: 51 932, 70 410) caribus em toda a região de estudo. A sub-área de Sisimiut apresentou o maior foco da população, seguida pelas sub-áreas Angujaartorfiup Nunaa e Sisimiut Sul, com estimativas de 46 724, 9 814 e 3 931 caribus, respetivamente. O DSM estimou uma abundância de cerca de 73 895 (IC 95%: 65 983, 82 757) caribus em toda a região Norte. É de notar que os intervalos de confiança se sobrepõem parcialmente, sugerindo que a diferença entre estas estimativas não é estatisticamente significativa. Em ambas as metodologias a

incerteza associada às estimativas obtidas foi relativamente baixa, sendo um indicador de precisão e consistência dos mesmos.

A diferença entre as estimativas de abundância pode ser justificada pelas características dos métodos aplicados. A amostragem por distâncias convencional é uma metodologia cuja extrapolação é *design-based*, o que significa que certas regiões ou características não amostradas ou mal representadas pelos transetos considerados não serão tidas em conta nas estimativas. Por outro lado, o GAM é uma abordagem *model-based*, considerando a informação sobre toda a região de estudo na modelação através das covariáveis ambientais fornecidas. Este segundo tipo de abordagem tem como vantagem a representação espacial da informação, permitindo assim o melhor entendimento e visualizção de possíveis padrões espaciais na abundância do caribu e, consequentemente, o afinamento das suas medidas de gestão. Outra vantagem adicional é permitir uma estimativa e respetiva incerteza associada em qualquer sub-área de interesse desde que esteja devidamente definida. Para ter uma estimativa da abundância, basta integrar sob a superfície de densidade estimada.

Por fim, comparando estas as estimativas pontuais com as obtidas em estudos anteriores, as primeiras são ligeiramente menores. Apesar disso, não existe evidência estatística que indique um decréscimo estatisticamente significativo na abundância dos caribus na região Norte do Oeste da Gronelândia. Recomenda-se, para estudos futuros e se possível, uma amostragem mais rigorosa no que diz respeito à categorização de covariáveis de interesse como a claridade, a profundidade e cobertura da neve, assim como a inclusão de outras covariáveis como a vegetação. Estas covariáveis poderão fornecer informação adicional sobre a distribuição da espécie de estudo.

Palavras-chave: caribu, densidade, Gronelândia, amostragem por distâncias, GAM, DSM

Abstract

The barren-ground caribou, *Rangifer tarandus groenlandicus*, is native to the west coast of Greenland, and has always been important for the human population. Its importance spans from cultural traditions and subsistence consumption to recreational and commercial harvesting. Hence, the importance of long-term monitoring to facilitate appropriate management strategies. To accomplish a robust monitoring method and to determine caribou density, Distance Sampling methods were used. These methods are widely used techniques for density and abundane estimation of a wide variety of taxa. In this project, the data from an aerial survey for caribou conducted by the Greenland Institute of Natural Resources (GINR) in the late winter of the year of 2018 was used to estimate abundance of caribou in the surveyed area. The survey data and the shapefiles with covariates to use in the DSM were provided by GINR. Starting from a conventional distance sampling perspective, the data set was then used to create a Density Surface Model (DSM) for caribou, *i.e.*, a model describing caribou density as a function of additional covariates collected during the survey. An introduction about the study species and the region of interest is provided as well as a brief description of the sampling design, Distance Sampling methodology and some related methods. The results were consistent with previous studies in terms of distribution throughout the study region, but the spatial distribution map obtained a previously unavailable useful insight. The estimated confidence intervals for abundance overlap with estimates from previous studies. Even though the point estimates are smaller when compared to previous point estimates, these differences are not statistically signifficant.

Key words: caribou, density, Greenland, distance sampling, generalized additive model, density surface modelling/spatial density modelling

Chapter 1

Introduction

This introductory chapter includes an overall description of the study species, as well as its habitat and the main purpose of the project.

1.1 The study species

The barren-ground caribou, *Rangifer tarandus groenlandicus* (Figure 1.1), is among the thirteen recognized subspecies of the genus *Rangifer* being one of those best studied within this taxa. This subspecies is native to Greenland and has been hunted for millennia. Since 1721, the caribou population has been known to fluctuate through cycles of exceptionally high and low abundance (Meldgaard, 1986), currently being in a period of high abundance (Cuyler et al., 2016).

Despite the severe inaccuracies in early surveys for caribou abundance, hunting was prohibited in 1993 and 1994, since managers were convinced that abundance was extremely low (Cuyler et al., 2007). According with some studies (Cuyler et al., 2003, 2007), local knowledge stated the opposite and was likely correct. Earlier monitoring resulted in little knowledge until the year 2000 as it produced unpublished studies and assessments. Only during the early years of this millennium have reliable monitoring methods (check Chapter 2) been applied for the surveys conducted by the Greenland Institute for Natural Resources (GINR) (Linnell et al., 2000; Poole et al., 2013).

Despite the socio-economic changes that have occurred in Greenland, the caribou is a species of both cultural and economic importance, but studies focused on caribou hunting are lacking, making it difficult to determine the value of the caribou in a social context (Linnell et al., 2000).

During the winter and early-spring period, when the aerial surveys occur, *R. t. groenlandicus* can usually be found in small groups consisting of four individuals (Table 5.1 in the Appendices), and they do not appear to aggregate into large coherent groups (Cuyler et al., 2005). Furthermore, they either present short migrations or do not migrate at all, which is unusual within the *groenlandicus* subspecies (Linnell et al., 2000). In fact, they have

Figure 1.1: *Rangifer tarandus groenlandicus*, the study species. A male calf (left) with polled cow (right), which trait is common in some Greenland populations. Photo provided by C. Cuyler.

complex and variable movement patterns, which complicates generalizing their behaviour. There are no natural predators where this subspecies is located (West Greenland). This allows hunting to be one of the main factors affecting caribou abundance (Linnell et al., 2000; Poole et al., 2013).

Two factors can act as abundance controllers or threats to the caribou. The first is density dependence. As a consequence of easily increasing their density, the caribou are exposed to increased intraspecific competition, which results may include reduced fecundity, increased calf mortality, and habitat overgrazing and trampling. Along with faeces contamination of the feeding area due to the high abundance of bacteria and parasites, other issues may arise such as diarrhoea and dehydration. The second factor that controls and/or threatens the caribou is the currently warming Arctic. Although the caribou is a circumpolar subspecies, its arctic habitat is very unique and highly vulnerable to climate change and global warming. This affects habitat characteristics and quality in several ways. Stochastic weather events can include deep snow, winter thaw-refreeze icing, rain-on-snow and summer droughts. These can compromise vegetation availability, quantity and quality. The extent, intensity and duration of such events influences caribou survival and productivity (Cuyler et al., 2002; Cuyler, 2007).

Human populations across the Arctic have traditionally relied heavily on natural resources, and *Rangifer* have always held a central role, being hunted for centuries in Greenland, as well as in Alaska, Canada, Russia and Northern Scandinavia. Given the caribou population growth in Greenland during the latter half of the 1900's (Cuyler, 2007), further research and reliable knowledge about this subspecies is needed regarding caribou abundance, spatial

Figure 1.2: A large group of the study species found within their habitat during an aerial survey. Photo provided by C. Cuyler.

distribution and location of essential areas, vital for survival and reproduction in order to adopt better management strategies and avoid adverse impacts on this valuable resource (Linnell et al., 2000; Tamstorf et al., 2005).

The barren-ground caribou play an essential cultural and economic role in Greenland, being both a commercial food resource and providing sport harvesting that maintains hunting traditions. Out-of-date and/or inaccurate information regarding the caribou population can result in management decisions that are inappropriate or even harmful to the caribou population sustainability (Cuyler et al., 2002, 2016). In a worst case scenario, this may lead to regional extirpation, as happened some places in Alaska during the 19th century, mainly due to human over-harvest (Linnell et al., 2000).

1.2 The study region

In West Greenland the majority of caribou occur within three large regions: the North region (Kangerlussuaq-Sisimiut population), the Central region (Akia-Maniitsoq population), and the South region (Ameralik and Qeqertarsuatsiaat populations) (Figure 1.3). This project is focused on the North region (Figure 1.4), which includes the city of Sisimiut and the town of Kangerlussuaq.

The North region, where the Arctic circle passes almost through the middle, covers an area

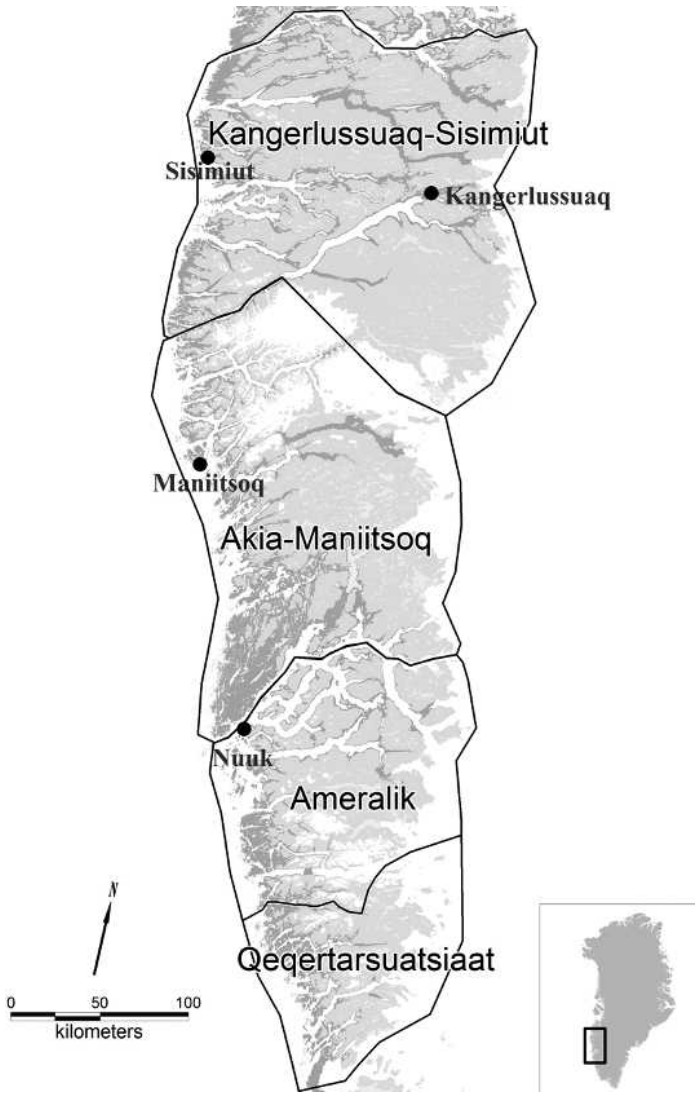

Figure 1.3: The four major caribou populations of West Greenland and basis for the aerial surveys performed by GINR (Poole et al., 2013).

of *circa* 24 000 km^2 seasonally ice-free, with a climate gradient, which is maritime at the seacoast but quickly becomes increasingly continental as one approaches the Greenland Ice Cap. In terms of topography, it is generally mountainous, open rough upland or alpine tundra (Figure 1.5). The Greenland Ice Cap forms the eastern boundary of the North region, while the southern border is formed by the much smaller Sukkertoppen Ice Cap, in combination with the outer Kangerlussuaq fjord, which is ice-free year-round. To the west is the Davis Strait, and the northern boundary is the Nordre Strømfjord, a maelstrom filled fjord, whose winter ice covering is incomplete, thin and treacherous.

The Kangerlussuaq-Sisimiut caribou population inhabits this region, and is the largest native population in West Greenland (Cuyler et al., 2011; Poole et al., 2013; Witting and Cuyler, 2008).

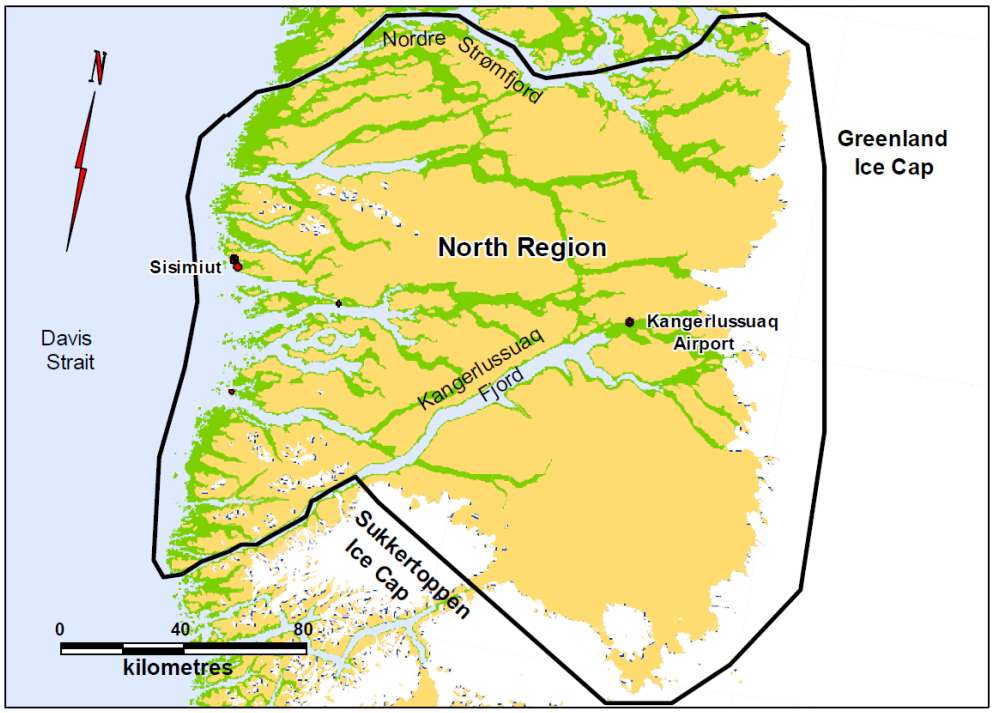

Figure 1.4: Main details of the North region of West Greenland (Cuyler et al., 2009).

Climate in the inland of the North region, is highly influenced by the Sukkertoppen Ice Cap, which is located south of the region. While acting as a barrier to incoming low-pressure storm systems, it creates a xeric continental climate, resulting in an annual precipitation around 150 mm and an annual mean July temperature of 11ºC (Tamstorf et al., 2005).

Vegetation is dominated by dwarf shrub heaths, grassland and steppe, such as *Empetrum nigrum* (crowberry) and *Vaccinium uliginosum* (bog bilberry) at low elevations in the coastal ranges, while *Betula nana* (dwarf birch) and *Salix glauca* (gray willow) dominate in the inland (Figure 1.6) (Tamstorf et al., 2005).

In the North region, grassland, willow and steppe comprise 70% of the vegetation types present, and these are somewhat positively related with a higher relative probability of use by caribou (Tamstorf et al., 2005).

Figure 1.5: Examples of the diversity of March snow cover present during aerial surveys of the North region, West Greenland. Photos provided by C. Cuyler.

Figure 1.6: Four of the main plant species present in the North region. *Empetrum nigrum* (crowberry) on the top left panel, *Vaccinium uliginosum* (bog billberry) on the top right, and *Betula nana* (dwarf birch) and *Salix glauca* (gray willow) on the bottom left and right panels, respectively (Tamstorf et al., 2005).

Given the characteristics of this region, its borders are efficient barriers, albeit semi-permeable, to mass caribou migrations between adjacent regions (Linnell et al., 2000).

1.3 The objectives of the project

Since the caribou have an important cultural and economic role in Greenland, management decisions will be facilitated by an improved understanding of relationships between caribou and the landscapes they inhabit. Therefore, the main purpose of this project was to estimate caribou density and abundance for the North region. We use data collected in March 2018 aerial surveys combined with Distance Sampling (DS) methodology. Then, create a Density Surface Model (DSM) for the caribou where its density can be spatially represented as a function of additional covariates collected during the surveys. The DSM is expected to produce further estimates of density and abundance for the North region and be an additional tool for caribou management in Greenland.

Chapter 2

Study description and design

Within the following chapter some considerations are made concerning the survey and its planning. First, the survey design and sampling are described, as well as the provided data set and software used during the analysis. This is followed by examples of previous studies and most relevant results observed through the years are presented for the study region. Finally, a brief description of the inherent Distance Sampling design and assumptions is made. Some of what is briefly presented within this chapter can be found in more detail in Chapter 3.

2.1 The survey design and sampling

Aerial surveys are performed using a helicopter. These are advantageous in many circumstances when compared to surveys performed by foot or handling a terrestrial vehicle. Large areas can be quickly surveyed in a short time and better visibility is provided, even through vegetation. Problems related with animal movement are often avoided since the helicopter speed is typically much greater than that of any target animal. However, flight speed also limits the time for detecting and identifying objects and recording data. Ideally, a good aerial survey should proceed at slow velocity with unrestricted forward and downward visibility, where speed and altitude depend on the characteristics of the object of interest (Buckland et al., 2001).

Additional data, such as group size or habitat-related features, may be useful to understand and describe the variables that affect density and distribution of the surveyed population. Other survey and environmental factors that affect visibility, such as flight altitude, solar glare and precipitation, or other cues that may affect detections can and should also be considered as covariates (Buckland et al., 2001).

The North region, including an area of about 24 000 km^2, was divided into three sub-areas, Sisimiut, Sisimiut-South and Angujaartorfiup Nunaa (Figure 2.1). Sisimiut is the largest with a total area of *circa* 13 000 km^2, while Sisimiut-South and Angujaartorfiup Nunaa are 3 500 km^2 and 7 000 km^2, respectively.

Concerning this project, and similar to Cuyler et al. (2011), the line transect was the selected sample unit for the survey thus, throughout this dissertation, only line transect design will be considered, even though there are other possible sampling methods (*e.g.*, point sampling).

The sampling covered a proportion of the survey region where the population of interest exists, and a certain number of sampling units, also called line transects, are defined *a priori* and distributed within this region. These transects correspond to rectangular shaped survey plots and the observer moves along the transect centre line. At the same time, the distances between the object of interest and the centre line are recorded (Buckland et al., 2001).

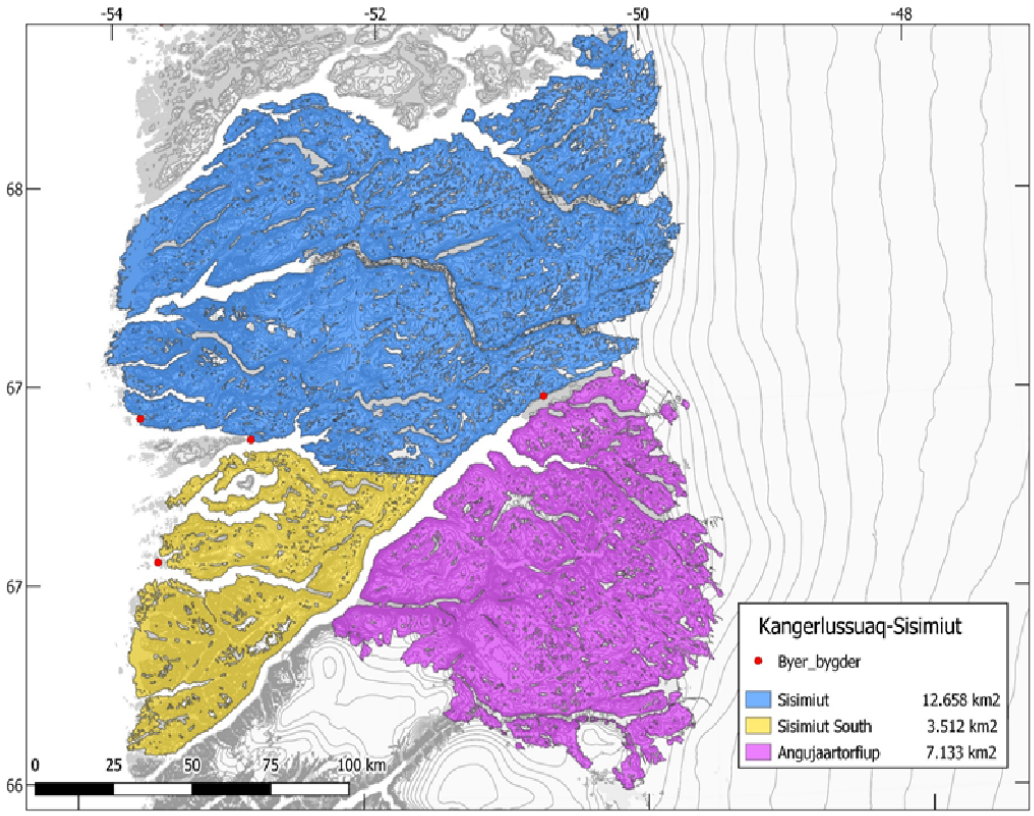

Figure 2.1: Sub-division of the northernmost survey region for the caribou aerial surveys. Each of the three sub-areas is represented by a different colour and their respective area provided. The term 'Byer bygder' identifies the main human settlements. Representation provided by C. Cuyler (2019).

For line transects, these distances are the perpendicular distances to the line and, from these recorded distances, a detection function can be estimated. From this, an estimate of the detection probability is obtained, and based on that, the density of the object of interest can be estimated for the surveyed area (Buckland et al., 2001).

Within the three sub-areas referred above, several line transects were defined (Figure 2.2). During each transect sampling, a helicopter with two main observers, one on each side, plus a third observer recording the observations, flew the transects during GINR's March 2018

aerial survey of the North region, West Greenland. For every detected group of caribou its perpendicular distance to the line and group size were recorded, as well as the observer that made the detection. Several covariates mostly related with the caribou and survey details, such as the Global Positioning System (GPS) coordinates, habitat and flight characteristics were also recorded. Other variables considered, because these could potentially explain caribou density, included aspect, elevation and slope. These were obtained in a Geographical Information System (GIS), details below.

The survey was conducted during March 2018. This was selected based on prior observations suggesting that this is the optimal survey timing, as observed by previous studies performed during the 1990s (Cuyler et al., 2016; Poole et al., 2013). For example, since the caribou dispersion is high and the group size is small and less variable, variance among transects is reduced and, as a consequence, counting error is diminished and precision is maximized. Further, in early March, caribou daily movement is at its minimum, which lowers movement between or along transects (Cuyler et al., 2016; Poole et al., 2013).

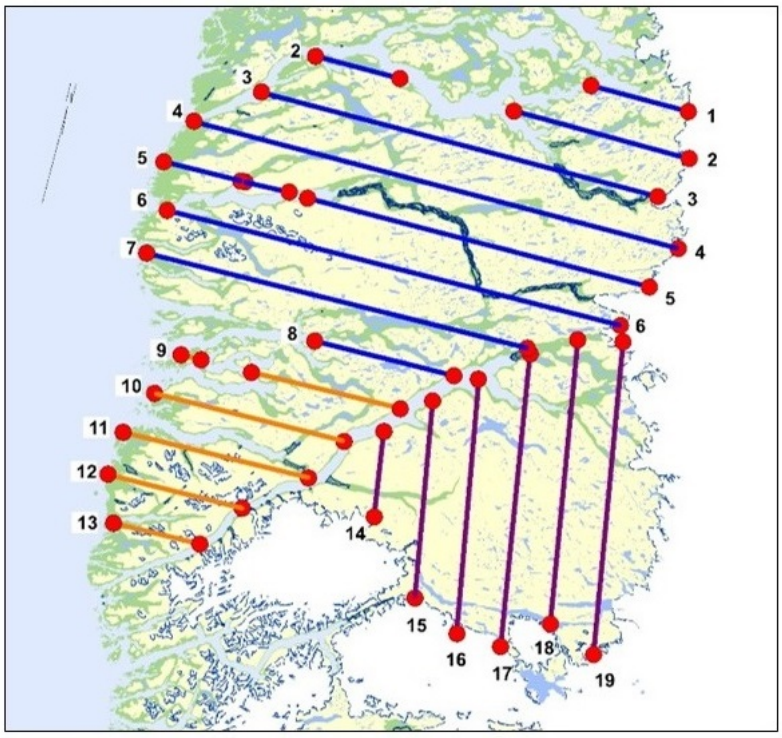

Figure 2.2: Line transects defined within each sub-area of the North region. Each of the three sub-areas is represented by a different colour: Sisimiut (blue), Sisimiut South (orange) and Angujaartorfiup Nunaa (purple). Representation provided by C. Cuyler (2019).

In total, 19 transects systematically spaced by 15 km were distributed throughout the three sub-areas (Figure 2.2). These transects provided the maximum coverage possible given the financial resources available. An initial transect was computationally generated at random in each strata, and others followed 15 km apart. The direction was chosen to be perpendicular

to previously known animal distribution gradients, which has been described to maximize precision by lowering encounter rate variance (*e.g.*, Buckland et al. (2001)).

2.2 Previous studies and results

Most surveys performed before 2000 had flaws crucial for the survey performance and were often invalidated by harvest data, *i.e.*, annual harvest similar to survey estimate yet population size increased over time (Cuyler, 2007). Flaws included observer fatigue due to the long survey duration, transect lengths of 80-100 km, high flight speed and high flight altitude, which disregarded Greenland's rugged terrain, and thus many caribou present remained undetected (Cuyler, 2007). The post-2000 surveys with their shorter, 7.5 km long transects, optimized sample size, variance, detectability, observer concentration and considered observer fatigue.

Cited through several reports and applied in previous studies, the survey characteristics before 2000 were very different when compared with those performed after the year 2000 (Table 2.1) (Cuyler et al., 2005, 2011). Clearly, several factors were not well managed or considered important in different steps of the pre-2000 surveys, so it is likely that those surveys underestimated population sizes, since most of the objects of interest were missed or not detected on the transects flown (Cuyler et al., 2005, 2011).

Table 2.1: Summary table comparing sampling design for aerial surveys before and after the year 2000, adapted from several references provided for the project.

Before 2000	After 2000
Long transects (80 to 100 km)	Short transects (7.5 km)
Long duration surveys (hours)	Short duration surveys (less than 1 h)
High speed flight (~167 km/h)	Low speed flight (45-65 km/h)
High and variable altitude (>152 m)	Low and constant altitude (15-50 m)
Flight direction from N to S	Change in flight direction
Sun glare unconsidered	Polarised sunglasses for sun glare
Without observer rotation	With observer rotation
Observer's fatigue unconsidered	Observer's fatigue considered
Wide strip width (1.4 km overall)	Short strip width (0.6 km overall)
Not tailored to rugged landscape	Adapted to rugged landscape
Unmarked windows	Marked windows for distance binning

Given the post-2000 changes to survey methods, precision increased and more animals present on the transects flown were actually detected. Consequently, the resulting estimates of population size far exceeded pre-2000 estimates, despite the lack of reported precision measures, such as standard deviance or confidence intervals (Figure 2.3).

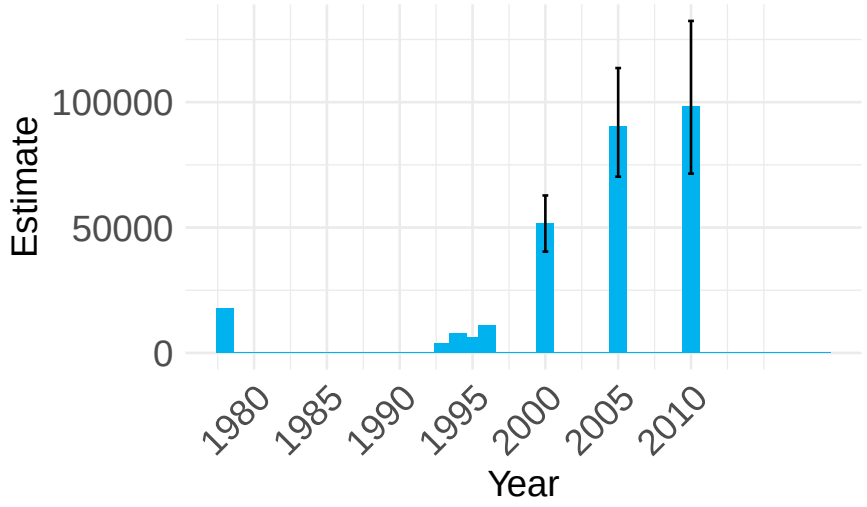

Figure 2.3: Population estimates for caribou in the North region. 90% confidence intervals are represented by black lines. Adapted from Cuyler et al. (2011).

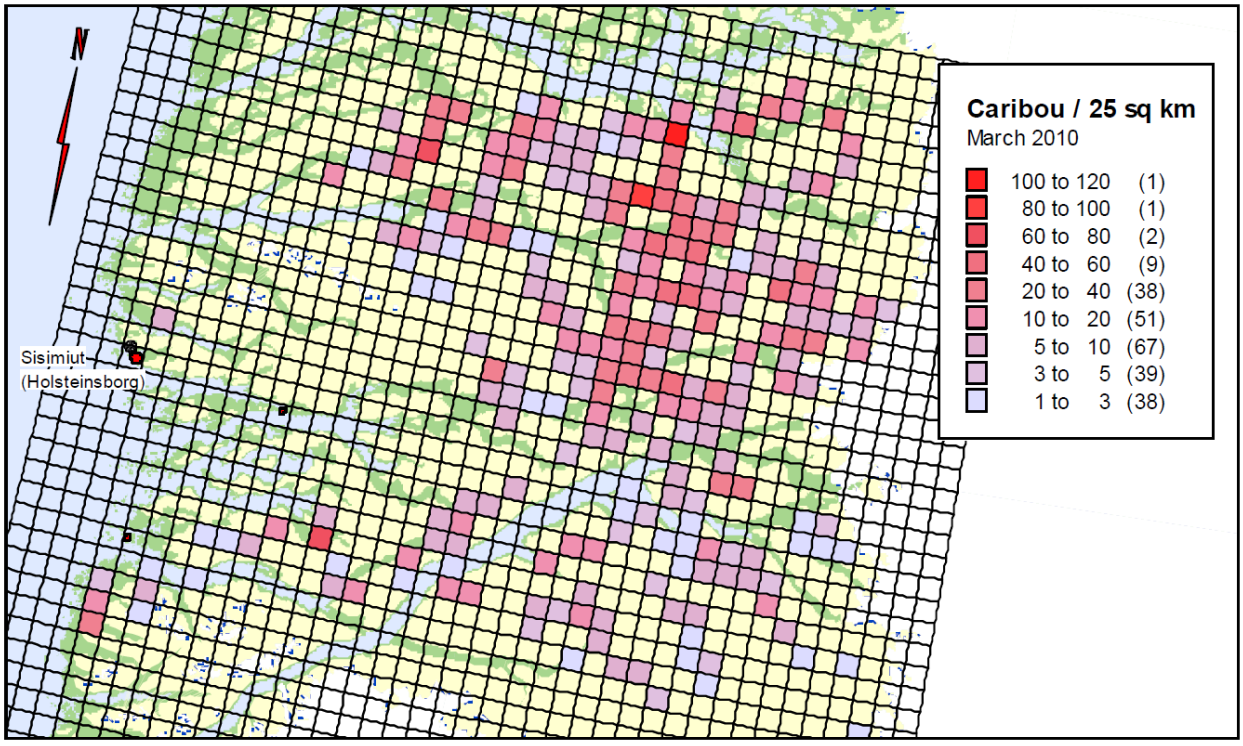

Figure 2.4: Heat map with the estimated density of the caribou throughout the North region obtained from 2010 survey. Coloured cells are those surveyed. Blank cells lack data and are not equivalent to a lack of caribou. Adapted from Cuyler et al. (2011).

The caribou in the North region increased during the 1990's and has been high since 2000 (Figure 2.3). The latter can lead to a population instability and, eventually, crash (Cuyler

et al., 2005). Since 2000, to control and reduce abundance, hunting season length was increased from one month to up to five and a half months. On the other hand, since 2010, season length varied being as short as three and a half to four months. Furthermore, since traditionally the majority of hunted animals were males, removing more females was also recommended to achieve these reductions (Cuyler et al., 2005; Cuyler, 2007). In 2019, the hunting season became a continuous five-month period (01 August - 31 December).

2.3 The data set and software

The data set for analysis was provided by C. Cuyler, Senior Scientist at the Department of Birds and Mammals from GINR. This data set included multiple variables and went through a cleaning process prior to the analysis.

Some other variables related with the environment were included in the analysis. These spatial variables, such as elevation, absent in the former data set, were sent in a shapefile format so these could be included in spatial modelling using GIS software. Chapter 4 has further details concerning the variables considered for data analysis.

All the statistical analyses were performed using R Statistical Software version 3.6.0. Shapefiles were manipulated in QGIS Software version 3.8.0.

2.3.1 Packages *Distance*, *knitr* and *bookdown*

Within the R Statistical Software, we used mostly the following three packages, **Distance** (Miller et al., 2016), **knitr** (Xie, 2015) and **bookdown** (Xie, 2020).

Distance was the main package that enabled the implementation of Distance Sampling analyses as it includes every functionality needed for data processing, model fitting, model selection and data reporting. This package is the "R-equivalent" of the standard Distance Sampling analysis software, DISTANCE (Thomas et al., 2010).

The two remaining packages are compilation packages that were used to write this dissertation. Whilst **knitr** allows dynamic report writing along with R programming coding and outputs, **bookdown** transposes these functionalities to the creation of larger documents, such as books, along with LaTeX and HTML coding flexibility for mathematical writing, text formatting and compilation.

2.4 Line Transect Distance Sampling assumptions and design

Line Transect Distance Sampling involves analyses where the primary variable is the distance from the line to the objects of interest. These can be animals, including those that typically

occur in clusters, plants or even inanimate objects such as nests, burrows, dung, whale blows or dead animals (Miller et al., 2016; Thomas et al., 2010).

Distance Sampling methodology is currently used to study many species around the globe (Couturier et al., 2018). For this project, some definitions must be first considered before entering into the statistical methodology. The distance measured is the perpendicular distance from line to the geometrical centre of the object of interest, in this case, groups of caribou.

Similar to any other statistical methods, Distance Sampling (Buckland et al., 1993), has some assumptions that must be met:

1. All objects directly on the line are detected;
2. Objects are detected at their initial location, independent on each other, and prior to any movement in response to the observer;
3. Distances (and cluster sizes, if applicable) are measured accurately.

Further assumptions are similar to other survey types, *e.g.*, each population is closed, being confined within a clearly defined area. Buckland et al. (2015) covers extensions that deal with the failure of some of these assumptions.

While selecting the survey design and respective line layout, five concepts should be addressed, and these are replication, randomization, sampling coverage, stratification and sampling geometry (Buckland et al., 2001).

Replication should be used in any survey design to provide a basis for an adequate variance of the encounter rate and a reasonable number of degrees of freedom for the confidence intervals estimation. It is recommended to consider a minimum of 10 to 20 replicate lines (Buckland et al., 2001). Some form of random probability sampling coverage or a systematic grid of lines with a random starting position is sufficient as long as the detections are independent and the transect systematic spacing and direction do not coincide with a regular spatial feature, resulting in an unrepresentative sampling (Thomas et al., 2010). Stratification, *i.e.*, partitioning the study region into smaller and homogenous areas before sampling, is important, if applicable. Finally, an adequate sampling geometry, which is related with the lines' configuration and orientation, should also be considered. These features depend on several factors including the survey region, logistics, efficiency, knowledge of density gradients or patterns and their relation with other sampling concepts.

Distances can be measured visually, directly, or with other devices such as binoculars or GPS devices. They must be accurately determined and may be grouped when exact measurements are difficult. In this situation, the distances might be recorded in intervals (bins), which is common in aerial surveys, using markers attached to the aircraft to delimit the distance intervals on the ground for a known flight altitude (Buckland et al., 2001).

According to Buckland et al. (2001), bias in distance measurements will result in bias in the density estimates. Visual estimation is the least favourable method and most prone to bias and errors, it should only be used as a last resort and observers must be well trained and tested (Anderson et al., 2001).

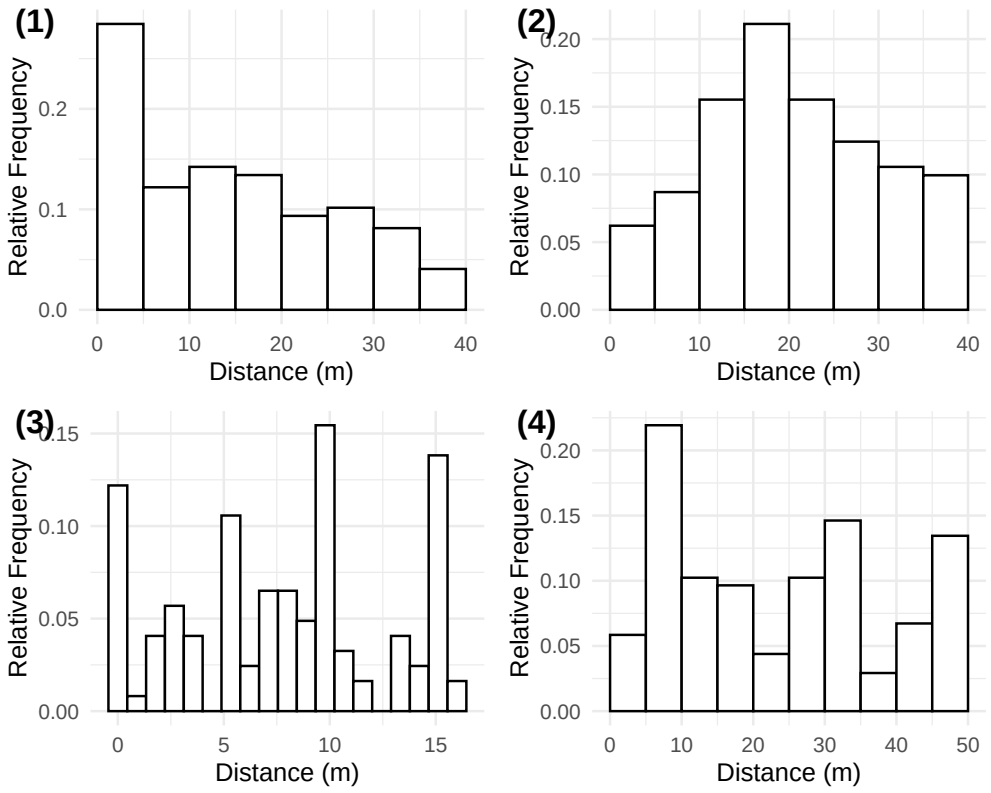

Figure 2.5: Examples of problematic line transect data sets: (1) spike at zero, (2) too few detections near zero, (3) rounding to favoured distances, (4) over-dispersed data. Adapted from Thomas et al. (2010).

With significant measurement error, four problems may arise: (1) a data spike near the centre line, (2) too few detections near the centre line, (3) rounding to favoured distances and (4) over-dispersed data (Figure 2.5). Spiked data, may arise when there are many detections near the line in comparison with all other distances. For example, in surveys where the study species is more detectable closer to the line and well camouflaged from afar. Too few detections near the centre line may occur in aerial surveys where it is difficult for observers to see the animals close to it. A possible solution for this situation includes aircraft with bubble windows, allowing the line to be seen. Another possible cause is animal movement away from the line prior to detection. In this case, attempts should be made to detect animals sooner, *e.g.*, by searching ahead in aerial surveys.

Rounding estimated distances to convenient values, such as multiples of 5 or 10, can cause considerable variability in the frequency counts since distances such as 1, 4 and 31 are less common (Buckland et al., 2001). Given sufficient data, rounding does not usually compromise estimation unless it is excessive. Better observer training can usually minimize this problem, also called *heaping* (Buckland et al., 2001). Another possible solution for heaping, is grouping distances into larger intervals that encompass the values at which heaping occurs (Buckland et al., 2015). Finally, over-dispersion may occur if animals appear in clusters, but are recorded as individuals. Superficially, over-dispersion looks similar to heaping (Figure 2.5 plots (3), (4)). However, larger frequencies do not occur at any obvious values to which distances might be rounded. This happens when it is not easy to locate the cluster centre, or to detect all animals within the cluster. In such circumstances, a recommended field protocol is to record each detected animal separately. This violates the independence assumption, but estimation is remarkably robust to even gross violations of this assumption (Thomas et al., 2010).

Chapter 3

Statistical methodology

Through this chapter the focus is set in the statistical theory behind the Distance Sampling methodology. At first, basic building blocks are presented in order to build the reasoning behind these methods. Secondly, the methods are described in a somewhat detailed way. At the end, descriptions of the Generalized Additive Model (GAM) and Density Surface Model (DSM) are presented. GAM and DSM are widely used approaches in this context and allow the connection between Distance Sampling estimates to an estimated spatial representation of density throughout the study region.

3.1 Distance Sampling

3.1.1 Fundamental concepts

Before entering into the detailed theory behind the Distance Sampling methodology, we present a simpler design, which is quadrat or plot sampling (Buckland et al., 2001; Marques, 2009).

In plot sampling, a region of interest with total area A, is divided into small plots of area a_{plot} (Figure 3.1). Some of these small plots are randomly chosen for sampling and the total number of individuals within these, n_{plot}, is recorded.

The density within each plot, D_{plot}, is the number of individuals per unit area for the respective plot so, by definition, it is given by

$$\widehat{D}_{plot} = \frac{n_{plot}}{a}, \tag{3.1}$$

where a is the total area sampled within A. (*i.e.*, $a = 4 \cdot a_{plot} = 4km^2$ for Figure 3.1). Since a random design was used, the density is a representative estimate, by design, for the total area A. Hence, an estimate for the abundance, $\widehat{N}$, can be obtained by simply multiplying $\widehat{D}_{plot}$ by the total area A,

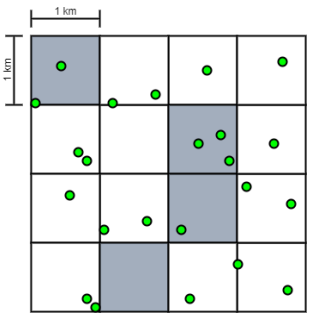

Figure 3.1: Plot sampling grid example of total area A divided into smaller plots of area a_{plot}.

$$\widehat{N} = A \cdot \widehat{D}_{plot} = A \cdot \frac{n_{plot}}{a}. \tag{3.2}$$

The DS methodology is an extension of quadrat-based sampling methods. The detail that creates the bridge from one methodology to the other is the fact that the method described above assumes that every individual of interest is detected (Miller et al., 2016). Frequently, this assumption cannot be met, specifically if among the individuals of interest there are animals impossible to observe owing to low sightability. Several factors cause low sightability, including topographical barriers, weather conditions, ground surface conditions and many others related to observer training and survey design. The proportion of individuals that were not detected can be estimated using the detection function fitted to the observed distances (Thomas et al., 2002). Once this proportion is estimated, it can be taken into account to obtain more accurate estimates and then, an extrapolation for a wider region can be done similarly as shown in Equation (3.2).

In Distance Sampling, this proportion of detected objects in the area a is defined as the probability of detection, P_a. Therefore, a density estimate can be obtained as per Equation (3.1) by adjusting n_{plot} by P_a, *i.e.*, by correcting the detections for those that were missed. Since the latter cannot be known, in general, an estimate must be also obtained, thus

$$\widehat{D} = \frac{\frac{n_{plot}}{\widehat{P}_a}}{a} = \frac{n_{plot}}{2wL\widehat{P}_a}, \tag{3.3}$$

where $\widehat{P}_a$ is an estimate of P_a obtained from the distance data, and a is the area of the sampled region. Usually $a = 2wL$, with w as the truncation distance, for both sides of the centre line, and the total transect length $L = \sum_{j=1}^{k} l_j$, where l_j is the length of transect j. Abundance can be determined using a reasoning analogous to that above (Equation (3.2)). The truncation distance is defined as the distance beyond which distances are not recorded. This can be defined in the field or at the analysis stage.

The coefficient of variation of $\widehat{D}$, $cv(\widehat{D})$, is related with two random components referred above, encounter rate (n_{plot}/L), and $\widehat{P}_a$, plus a third one that is the estimate of the expected

size of detected clusters ($\widehat{E}(s)$). Assuming independence between these, the former is given by

$$(cv(\widehat{D}))^2 = \left(\frac{se(\widehat{D})}{\widehat{D}}\right)^2 = (cv(n_{plot}/L))^2 + (cv(\widehat{E}(s)))^2 + (cv(\widehat{P}_a))^2. \tag{3.4}$$

An approximation of the standard error of $\widehat{D}$, $se(\widehat{D})$, is defined as

$$se(\widehat{D}) = \widehat{D} \cdot \sqrt{(cv(n_{plot}/L))^2 + (cv(\widehat{E}(s)))^2 + (cv(\widehat{P}_a))^2}. \tag{3.5}$$

Once these are obtained, an approximate $100(1-\alpha)\%$ confidence interval (CI) can be determined by

$$\widehat{D} \pm z_{1-\frac{\alpha}{2}} \cdot se(\widehat{D}), \tag{3.6}$$

where $z_{1-\frac{\alpha}{2}}$ is the quantile of the $N(0,1)$ distribution ($z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96$ for a 95% confidence interval). However, the distribution of $\widehat{D}$ is positively skewed, thus an interval assuming that $\widehat{D}$ is log-normally distributed has better coverage. According with Buckland et al. (2015), a $100(1-\alpha)\%$ confidence interval can be given by

$$\left(\widehat{D}/C, \widehat{D} \cdot C\right), \tag{3.7}$$

where

$$C = \exp\left\{z_{1-\frac{\alpha}{2}} \cdot se[\log_e(\widehat{D})]\right\} \tag{3.8}$$

and

$$se[\log_e(\widehat{D})] = \sqrt{\log_e\left[1 + (cv(\widehat{D}))^2\right]}. \tag{3.9}$$

For further details see Buckland et al. (2001) and Buckland et al. (2015).

3.1.2 Probability of detection

Given the above, the probability of detecting an object, giving that it is within the area covered by the transects, $\widehat{P}_a$, needs to be estimated. For this project, the object of interest consists in caribou groups.

To illustrate the importance of this probability, consider that an observer is walking across a large patch of tundra and detects 8 caribou. While discussing with the local biologist, and

considering the biologist's experience, he/she will state that, on average, only one third of the caribou are detected (*i.e.*, $\widehat{P}_a = 1/3$) meaning that probably there were around 24 caribou within the tundra and 16 have been missed. That is where Distance Sampling is useful, since it allows a rigorous framework for the estimation of P_a and then an estimate of abundance can be obtained as shown in Equation (3.3).

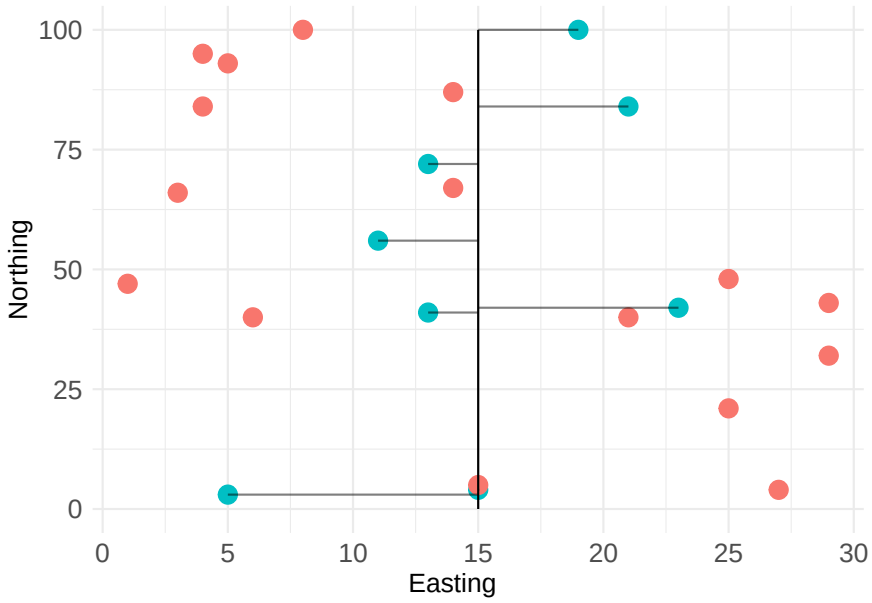

Figure 3.2: Example of a patch of tundra with the transect in the middle. Blue dots represent observed caribou, while orange dots represent the undetected ones. The lines perpendicular to the transect represent the recorded distances.

3.1.3 Distance Sampling methods

The detection function, $g(y)$, describes the probability of detecting an object of interest given that it is at a distance y, from the centre line, thus being a non-increasing function of y (Buckland et al., 2015).

For line transects, y is the perpendicular distance from the 0-line to the detected object. Within Distance Sampling methods, the probability of detection is explained recurring to these observed distances (Buckland et al., 2001). Sometimes covariates may be added to explain their relation with the detection probability. In this situation, we are within the Multiple Covariate Distance Sampling (MCDS) framework (Buckland et al., 2001).

Conventional Distance Sampling

Conventional Distance Sampling (CDS) occurs when no additional covariates are added to the model. Once the detection function is estimated, $\widehat{P}_a$ can be obtained via the following equation

$$\widehat{P}_a = \int_0^w \hat{g}(y) \cdot \pi(y) dy, \tag{3.10}$$

where $\pi(y) = \frac{1}{w}$ and, therefore, used to estimate density using Equation (3.3). For $g(y)$ it is also specified a flexible semi-parametric model, composed by a key function and some additional series expansions, known as adjustment terms, and their parameters are estimated (Marques et al., 2007).

In order to obtain robust estimates of density, flexible models for $g(y)$ are needed with the form (Buckland et al., 2001)

$$g(y) = \frac{k(y) \cdot [1 + s(y)]}{k(0) \cdot [1 + s(0)]}, \tag{3.11}$$

where $k(y)$ is the parametric key function and $s(y)$ represents the additional adjustment terms (Table 3.1).

Table 3.1: Commonly used key functions and series expansions for the detection function. Adapted from Buckland et al. (2001).

Key function		Series expansion	
Uniform	$1/w$	Cosine	$\sum_{m=2}^{M} a_m cos(m\pi y_s)$
Half-normal	$\exp[-y^2/2\sigma^2]$	Simple Polynomial	$\sum_{m=2}^{M} a_m (y_s)^{2m}$
Hazard-rate	$1 - \exp[-(y/\sigma)^{-b}]$	Hermite	$\sum_{m=2}^{M} a_m H_{2m}(y_s)$

Note: If Uniform key, $m = 1, ..., M$. $H(x)$ denotes Hermite function.

The uniform key function has no parameters, while the half-normal and the hazard-rate functions include a scale parameter, σ, which determines the rate at which the function decreases with increasing distance (Figure 5.1). Furthermore, the hazard-rate function also includes a shape parameter, b, that provides greater flexibility to this function comparing to the others (Buckland et al., 2001).

It is not always necessary to include adjustment terms and, in such cases, these models are referred to as "key only" models. When the key functions are not enough for fitting $g(y)$, some series expansions terms may be added to modify its shape (Figure 5.2). These terms can be either cosine, simple polynomial or Hermite polynomial (Table 3.1).

It is important to note that these adjustment terms do not depend directly on y but on y_s, which is a scaled value of y, where $y_s = \frac{y}{w}$ with w being the truncation distance. This allows independence between the shape of the series expansion and the units used for y (Marques et al., 2007).

Right truncation of the data, or the removal of the largest distances, is a common procedure that aids model fitting. Some precision might be lost with truncation, however it is usually slight. On the other hand, precision is increased since the data is easier to model and,

consequently, fewer parameters and adjustment terms are required to model the detection function (Couturier et al., 2018).

Multiple Covariate Distance Sampling

CDS methods can be extended to MCDS, so that $g(y)$ is modelled as a function not only of distance, but also of a vector of J additional covariates for each of the n objects of interest, $\mathbf{z_i} = z_{i1}, ..., z_{iJ},\ i = 1, ..., n$. Accordingly, the function that describes the probability of detection at a given distance, is represented by $g(y, \mathbf{z})$. These additional covariates can either be discrete or continuous, such as observer and group size, and are assumed to affect only the scale, σ, of the detection function (Marques et al., 2007; Miller et al., 2016). For line transects, $P(\mathbf{z_i})$, *i.e.*, the probability of detecting the i-th object of interest given its respective vector of covariates $\mathbf{z_i}$ can be estimated using the formula presented in Equation (3.12).

$$\widehat{P}(\mathbf{z_i}) = \int_0^w \hat{g}(y, \mathbf{z_i}) \cdot \pi(y) dy, \tag{3.12}$$

with $\pi(y) = \frac{1}{w}$. Considering the three key functions previously presented, only the uniform key is excluded from MCDS since it does not have a scale parameter. Half-normal and hazard-rate functions can have their scale parameter written as a function of the covariate values as

$$\sigma(\mathbf{z_i}) = \exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j z_{ij}\right), \tag{3.13}$$

where β_0 and all the β_j's are the $J + 1$ coefficients to be estimated with J being the total number of covariates. The estimation of the parameters for both CDS and MCDS is typically done via maximum likelihood (Marques et al., 2007).

Once the detection function is estimated, according with (Buckland et al., 2004), density can be estimated as

$$\widehat{D} = \frac{1}{a} \sum_{i=1}^{n} \frac{1}{\widehat{P}(\mathbf{z_i})}, \tag{3.14}$$

where a is the total area surveyed, $\widehat{P}(\mathbf{z_i})$ is the estimated probability of detecting the i-th object of interest given its respective vector of covariates $\mathbf{z_i}$.

Finally, Marques et al. (2007) states that MCDS methods potentially offer improved inference in four situations, when comparing to CDS methods:

1. when a subset of data is used to estimate density, *e.g.*, by strata, where this information can be introduced as a factor covariate. In CDS, the strategy is more complex, either

to estimate P_a for each stratum and thus, stratum-level estimates for density or to use a global estimate for the probability of detection, but this second introduces bias, for example, if one stratum favours the animals when compared to other strata which uses fewer parameters than a fully stratified detection function model;

2. where pooling robustness does not hold for CDS analyses, *e.g.*, when survey intensity varies according with pre-defined strata to increase efficiency, or when the detection probability faces extreme heterogeneity due to different object habitats or behaviours, for example, showy males contrasting with cryptic females in animal surveys;

3. reduces the variance of density estimates by modelling the heterogeneity in the detection function;

4. if there are covariates of interest to be included in the model.

3.1.4 Model selection

Since the estimator of density is closely linked to the detection function, it is of critical importance to select models for the detection function carefully. Three properties desired for a model for $g(y)$ are, in order of importance, *model robustness*, a *shape criterion* and *estimator efficiency* (Buckland et al., 2001, 2015; Miller et al., 2016).

The most important property of a model for the detection function is model robustness. According with Buckland et al. (2001) and Buckland et al. (2015), this means that the model is a general, flexible function that can take a variety of plausible shapes for the detection function. The concept of pooling robustness is also included here. Models of $g(y)$ are pooling robust if the data can be pooled over many factors that affect detection probability and still yield a reliable estimate of density. A model is pooling robust if, for example, a stratified estimation for density, $\widehat{D}_{st}$, and a pooled estimation for density, $\widehat{D}_p$, are approximately the same. In the first scenario, the data is stratified by factors, such as observer or habitat type, and an estimate for density in each stratum is made. Then these estimates are combined into $\widehat{D}_{av}$, an average density estimate. In the second scenario, all data could be pooled, regardless of any stratification, and a single estimate computed, $\widehat{D}_p$. A model is pooling robust if $\widehat{D}_{av} \approx \widehat{D}_p$.

According to Buckland et al. (2001), the shape criterion consists in the fact that the detection function should have a 'shoulder' near the line (Figure 3.3), *i.e.*, detection remains nearly certain at small distances from the sampling unit's centre line ($g'(0) = 0$). This allows the reliable estimation of object density (Thomas et al., 2002). Generally, good models for $g(y)$ will satisfy the shape criterion near the zero distance line, which is especially important in the analysis of data where some heaping at zero distance is suspected.

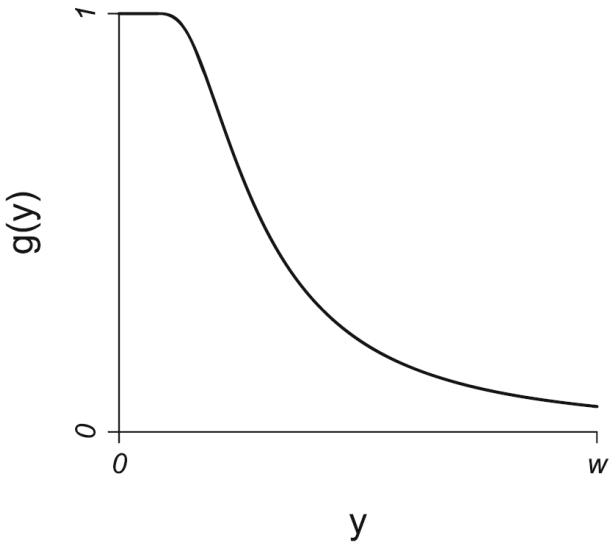

Figure 3.3: A good model for the detection function should have a shoulder, with probability of detection staying at or close to one at small distances from the line or point. At larger distances, it should fall away smoothly. The truncation distance w corresponds to the strip half-width (for Line Transect Distance Sampling). Adapted from Buckland et al. (2001).

Estimator efficiency is the third most important property (Buckland et al., 2001), which means that it is desirable to select a model that provides estimates that are relatively precise, *i.e.*, that have small variance. This property is of benefit only for models that are model robust and have a shoulder near zero distance, otherwise the estimation might be precise but biased.

Besides these three criteria, the model should be a monotonic function of distance from the line, that is, the probability of detection at a given distance cannot be greater than the probability of detection at any smaller distance (Figure 3.3) (Buckland et al., 2001).

There is no fixed standard method to select the best fitting model, *i.e.*, choosing the most appropriate key function and series expansion (Marques et al., 2007). It is usually done by applying the Akaike's Information Criterion (AIC), Kolmogorov-Smirnov test, Cramér-von Mises test and the χ^2 Goodness-of-Fit test (GOF test). The likelihood ratio test can also be used but, since it is only applicable for nested models, AIC is the recommended method (Marques et al., 2007). A proper model should be simple with an adequate fit without overfitting the data.

Akaike Information Criterion

The relative fit of alternative models may be evaluated recurring to AIC, or AICc, in case of small samples, providing a small sample bias correction (Buckland et al., 2001). These criteria can be determined as follows

$$AIC = -2 \cdot \ln(\mathcal{L}) + 2q, \text{ and} \tag{3.15}$$

$$AICc = AIC + \frac{2q(q+1)}{n-q-1}, \tag{3.16}$$

where $\mathcal{L}$ is the likelihood function, q is the number of estimated parameters in the model, and n is the sample size. This measure provides a trade-off between bias and variance. AIC includes two terms, one related with the fitted model, and the other working as a penalty considering the excess of parameters in the model (Brewer et al., 2016).

Kolmogorov-Smirnov test

The Kolmogorov–Smirnov test is one of the tests that can be applied to the detection function to assess model fit (Buckland et al., 2004). This test is only applicable for continuous data, being preferable to the χ^2 GOF test for MCDS methods.

Considering the cumulative distribution function (c.d.f.) $F(x) = P(X \leq x)$ and the empirical c.d.f. (e.d.f.) $S(x)$, the null hypothesis to be tested is $H_0 : F(x) = F_0(x), \forall x$. The alternative hypothesis states that both functions differ for at least some value of x. In practice, $F(x)$ is replaced by its estimate, and H_0 states that the assumed model is the true model for the data (Buckland et al., 2004). The largest absolute difference between $\widehat{F}(x)$ and $S(x)$, denoted D_n, is the test statistic (Gibbons and Chakraborti, 2011). The corresponding p-value can be approximated by

$$p = 2 \cdot \sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2ni^2 D_n^2). \tag{3.17}$$

Cramér-von Mises test

Similarly to the Kolmogorov-Smirnov test, the Cramér-von Mises test shares the same null hypothesis and basis on diferences between c.d.f. and e.d.f.. However, instead of considering only the largest difference between the two functions, this test is based on their entire range (Buckland et al., 2004). The test statistic can be given by

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\widehat{F}(x_{(i)}) - \frac{i-0.5}{n} \right]^2. \tag{3.18}$$

Chi-square Goodness-of-Fit test

The χ^2 Goodness-of-Fit test (Buckland et al., 2001, 2015) compares the observed frequencies, n_i, with the expected frequencies under the model $E(n_i)$, and it is given by:

$$X^2_{obs} = \sum_{i=1}^{n} \frac{\left[n_i - E(n_i) \right]^2}{E(n_i)} \sim \chi^2_{(u-q-1)}, \tag{3.19}$$

under the null hypothesis (H_0) of good model fitting, *i.e.*, the difference between the observed (n_i) and expected ($E(n_i)$) counts is close to 0. In Equation (3.19), n is the total number of observations, u is the number of groups (or bins) within the distance data, and q is the number of model parameters estimated. Reject H_0 if $\chi^2_{obs} > \chi^2_{1-\alpha;(u-q-1)}$, with the latter representing the $1-\alpha$ quantile from a χ^2 distribution with $u-q-1$ degrees of freedom.

As the number of parameters of the fitted model increases, the bias decreases, but the sampling variance increases (Buckland et al., 2001). While the Goodness-of-Fit test results should be considered in the analysis of distance data, they will be of limited value in selecting a model since these tests are sensitive to heaping. Therefore, care is needed in choosing suitable distance intervals.

If data are collected with no fixed w, it is possible that a few extreme outliers will be recorded. These values are not useful and the data should therefore be truncated. This can be checked using the distances' histogram, and also whether there is evidence of heaping or not (Buckland et al., 2001; Couturier et al., 2018).

Goodness-of-Fit tests allow formal testing of whether a detection function model provides an adequate fit to the data. Since the GOF test cannot be used on continuous data, unless grouped, it is of limited use for testing MCDS models (Buckland et al., 2015), being useful for testing models using CDS methods. However, if distances are not grouped, they must first be categorized into groups to allow the test to be conducted. Thus there is a subjective aspect to the test, and different analysts, using different group cutpoints, may reach different conclusions about the model adequacy. In contrast, the Kolmogorov–Smirnov and Cramér–von Mises tests can only be applied to continuous data (Buckland et al., 2015).

3.2 Generalized Additive Models (GAM)

Generalized Additive Models are an extension to Linear Models (LM) and Generalized Linear Models (GLM) where non-linear responses with smoothing functions can be fitted to the data. So, to define a GAM, firstly LM and GLM will be briefly introduced.

In LM, and considering a total of n observations, a response variable Y_i, with mean value $\mu_i = E(Y_i)$ is expressed as the sum of a function of linear combinations of M independent variables X_m, $m = 1, 2, ..., M$, and a random error ϵ,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_M X_{Mi} + \epsilon_i, \quad i = 1, ..., n, \tag{3.20}$$

where the β's are the unknown parameters, with β_0 representing the intercept and β_M the parameter associated with the M-th predictor. It is assumed that the random errors ϵ are independent and identically distributed, following a $N(0, \sigma)$ distribution (Wood, 2017).

GLMs are an extension of LM where there can be a non-linear relationship on the response scale between the dependent variable and the independent variables (Wood, 2017), considering that the former follows a distribution from the Exponential family and using a link function, $g(.)$. Therefore, the equation of the model is given by

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_M X_{Mi}. \tag{3.21}$$

Finally, GAM are a GLM extension where the linear predictor includes smoothing functions of covariates, $f(.)$ (Wood, 2017). As happens with GLM, Y_i must belong to the Exponential family and a link function, $g(.)$, is specified. A GAM is expressed as

$$g(E(Y_i)) = \beta_0 + f_1(X_{1i}) + f_2(X_{2i}) + f_3(X_{3i}) + ... + f_M(X_{Mi}). \tag{3.22}$$

These smoothing functions need, therefore, to be defined and predetermine how flexible or smooth they should be. Another advantage of working with this type of functions is the possibility of combining covariates into a single smoothing function. Thus, by combining, for example, the covariates X_1 and X_2, the above formulation would look like

$$g(E(Y_i)) = \beta_0 + f_1(X_{1i}, X_{2i}) + f_3(X_{3i}) + ... + f_M(X_{Mi}). \tag{3.23}$$

Therefore, GAM are more adequate since these are able to fit smoothing functions through the data, which provides flexibility in describing complex non-linear relationships (Wood, 2017).

Furthermore, if Y_i is described by a Poisson distribution, the surveyed area is defined as the offset variable. This variable can be added to the model with its regression coefficient known to be 1. Equation (3.22) can then be rewritten as

$$g(E(Y_i)) = \beta_0 + f_1(X_{1i}) + f_2(X_{2i}) + f_3(X_{3i}) + ... + f_M(X_{Mi}) + \log(offset). \tag{3.24}$$

During this project, the offset term considered to explain the number of observations in a given area was the effective sampling area, obtained in the first stage via MCDS.

3.3 Density Surface Model

Conventional Distance Sampling methods provide average estimates of abundance over a region but no information about the distribution of the objects of interest within the survey region. One possible option is to divide the study area into progressively smaller and smaller strata to try to detect patterns in spatial distribution. However, a more efficient approach is to build a spatial model. These models incorporate spatially-referenced environmental covariates, thus modelling part of the spatial variability of the data (Katsanevakis, 2007; Miller et al., 2013). Density surface modelling uses the GAM framework (Wood, 2017) to build models of abundance/density as a function of environmental covariates, typically as part of a two stage method. In the first stage we model the detectability via distance sampling and in the second stage we model the counts, corrected for detectability, over space. To propagate the variance, the GAM model parameters' distribution can be used to generate

replicate abundance estimates, and the uncertainty from the estimated detection function can be incorporated as an additional random effect term (Miller et al., 2013).

Generally, very little information is lost by taking this two-step approach as the transects are very narrow when compared with the width of the study area. So, provided no significant density variation takes place across the lines' width or within the point, there is no information in the distances about the spatial distribution of animals (Miller et al., 2013).

Additionally, DSM can be used to predict abundance and density over an area of interest, given the other environmental covariates (Miller et al., 2013). In our work, the provided covariates of interest included aspect, elevation, slope and vegetation. Although relevant, the latter was discarded since available data lacked the necessary resolution for the analysis.

After fitting the DS model, the counts within the defined offset are adjusted with the estimated probability of detection and, with this new data set including the offset, a GAM is fitted. A prediction map can then be created once a spatially georeferenced data set used for prediction is obtained. Furthermore, it would be better to include the uncertainty associated with each mapped estimate, since these estimates alone can be less informative. Within this project, a map representing the coefficient of variation (CV) associated to each estimate was obtained resorting to the packages mentioned in the previous chapter, even though there are other possible methods (*e.g.*, varying each prediction's pixel size according with the associated uncertainty, within a single map).

Using QGIS, the whole study region was converted into small cells. Centered in each point, a $1.5km \times 1.5km$ square shaped buffer zone was created, and intersected with the sampling points and spatial covariates. For each square, the mean of each covariate was determined, along with the respective central pair of coordinates. This information was converted into a data set to create the GAM model.

Figure 3.4 represents a flow diagram of the whole process during the analysis, from the Distance Sampling analysis to the Density Surface Modelling.

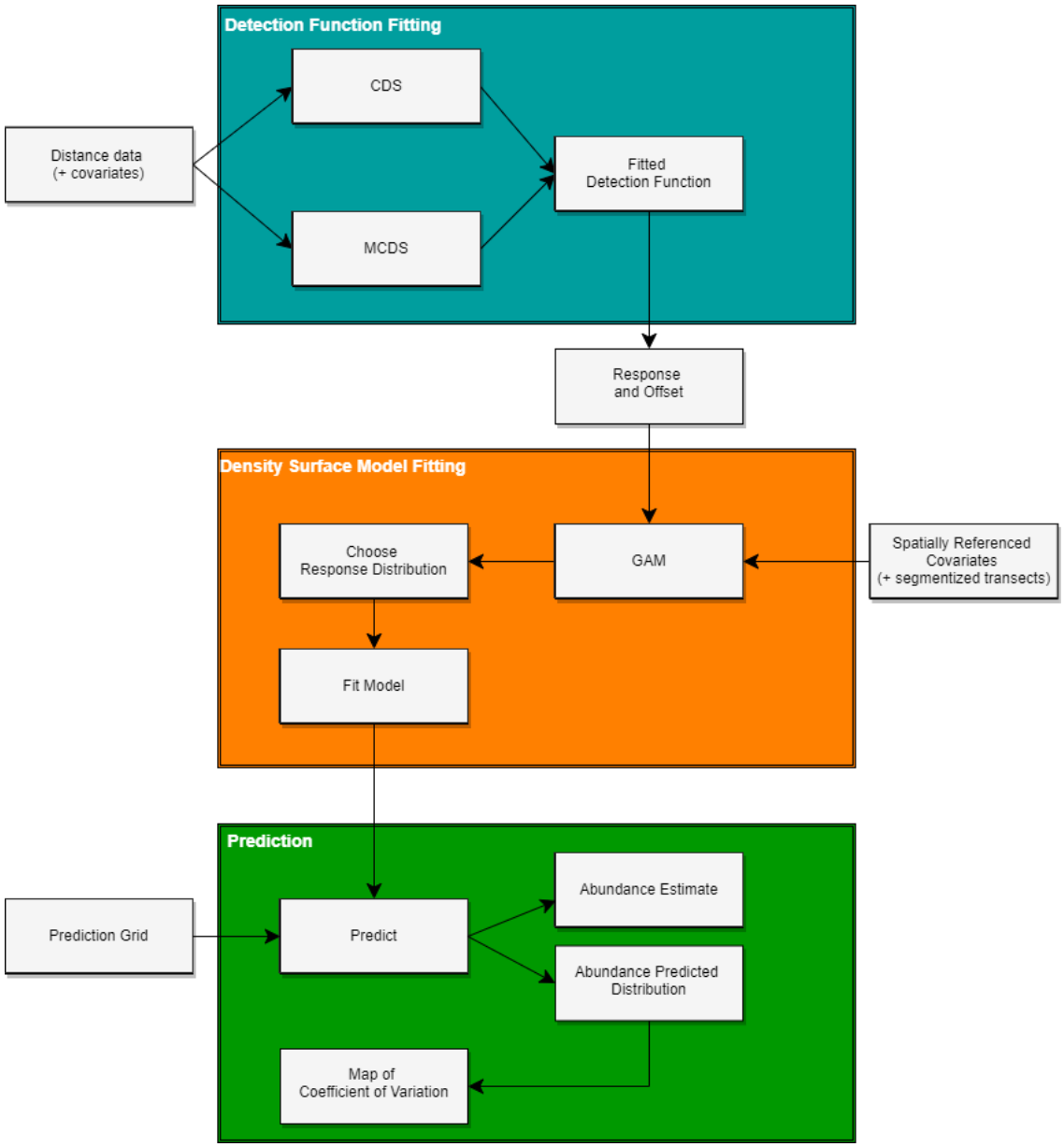

Figure 3.4: Flow diagram showing the modelling process for creating a Density Surface Model. Created using Draw.io version 12.7.9. Adapted from Miller et al. (2013).

Chapter 4

Statistical analysis and results

This chapter reports on the data analysis. Firstly, data preparation for analysis and exploratory analysis are presented. This is followed by DS model fitting and model selection. To finish, the GAM is fitted and the resulting DSM is presented.

4.1 Data provided and prior processing

The data set was provided in an Excel format with the survey variables, such as region, sub-area, respective areas (in km^2) and transect identification and characteristics, and was subject to some prior processing before being further analysed. This data set included variables such as recorded distances, group size, GPS coordinates, flight and survey characteristics such as glare, shade, snow covering and depth. Most of these were not included in the analysis due to inconsistent categorization and high incidence of missing observations. The remaining variables were properly restructured within R Statistical Software.

During the survey, the distances were recorded in bins of 0, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75 and 1.50 km. Bigger bins for larger distances are a common strategy in this type of analysis since larger distances are more prone to errors, therefore improving precision (Buckland et al., 2001). Then these distances were corrected for Distance Sampling analysis, more specifically, "mid-binned" because the distance values corresponded to the upper limit for the bin in which the observation was included. A total of 5060 caribou in 2076 groups were detected during the survey.

A small number (10) of observations were missing their distance. For these, the average observed distance was imputed, a pragmatic solution for the observations to be considered. For group size, only one observation had no value. As with missing distance, the average group size was inputed to this observation.

4.2 Exploratory analysis

As suggested by several authors (Buckland et al., 2001; Marques et al., 2011; Thomas et al., 2010), firstly an extensive exploratory data analysis must be made, then model fitting and selection. The following analysis is based on Marques (2018b) and Marques (2018a).

Buckland et al. (2001) suggest the sample size to be at least 60 to 80 observations and also a minimum of 10 to 20 replicate line transects to allow reliable estimates. Within this project, these suggestions were met, since we had a sample size of 2076 observations, *i.e.*, detections of groups of (one or more) caribou, and a total of 19 parallel line transects 15 km apart.

Each transect, depending on number of segments and total length, took from as few as 30 min to several hours to be fully sampled with a constant flight altitude around 40 m above the ground surface, thus following terrain's topography.

Concerning the number of detections, *i.e.*, caribou groups, per sub-area, one can clearly see that the sub-area of Sisimiut dominates in observation frequency, as expected, since it is the largest sub-area, with *circa* 13 000 km^2. The other two sub-areas, Sisimiut South and Angujaartorfiup Nunaa, having around 3 500 and 7 100 km^2, respectively, had fewer detections (Figure 4.1).

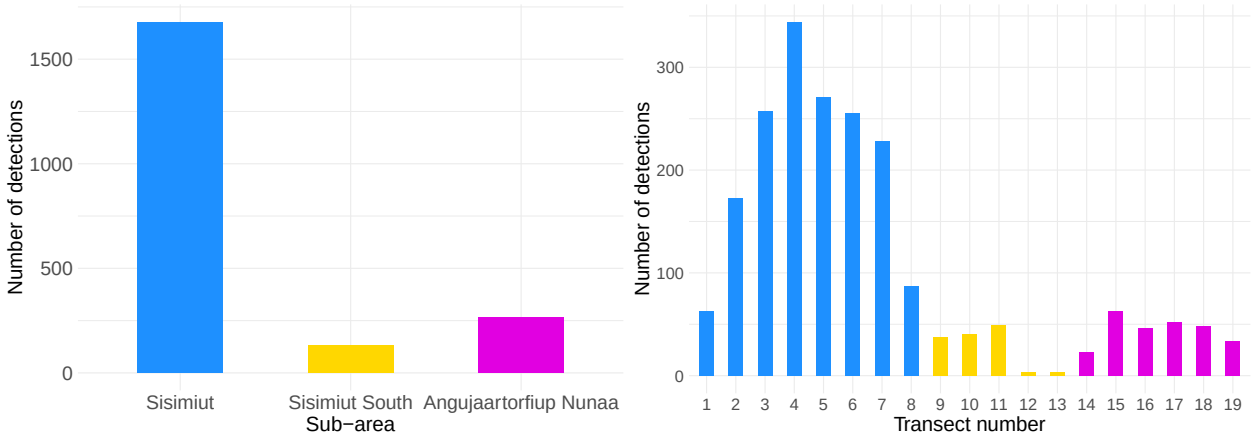

Figure 4.1: Exploratory analysis plots for the caribou survey data: number of detections by sub-area (left), and number of detections per line transect by sub-area (right) each with a different colour, Sisimiut (blue), Sisimiut South (yellow) and Angujaartorfiup Nunaa (magenta).

Furthermore, the caribou were detected on every line transect in the survey, with the Sisimiut sub-area distinguishing itself from the remaining with a dominating number of detections (Figure 4.1).

The objects of interest, *i.e.*, caribou groups, that were detected usually included no more than six animals (Figure 4.2). The most common observed group size was two individuals (733 observations). In total, around 90% and 99% of the observations consisted in groups with less than five and ten individuals, respectively. Larger groups were scarce and typically

observed at greater distances, for example, the largest caribou group size, observed twice, had 20 caribou and both were detected at 1.1 km from the line.

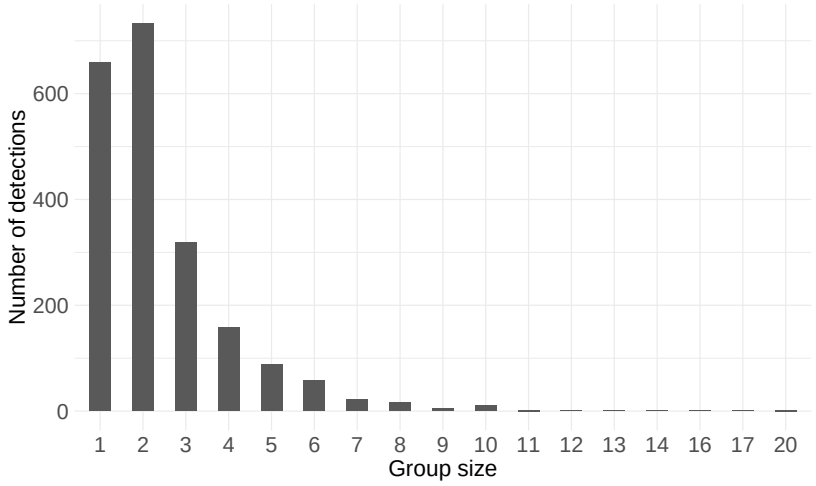

Figure 4.2: Exploratory analysis plot for the caribou survey data: group size distribution.

Most transects were surveyed either from East to West or from West to East. However, the transects within the Angujaartorfiup Nunaa sub-area were surveyed from North to South and South to North directions, given the climate and animal distribution gradients.

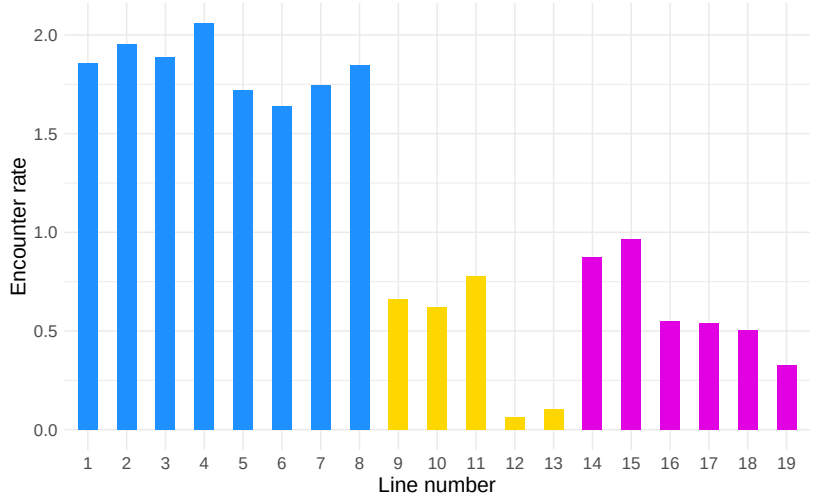

Figure 4.3: Exploratory analysis plot for the caribou survey data: caribou encounter rate per transect by sub-area in groups per km, each with a different colour, Sisimiut (blue), Sisimiut South (yellow) and Angujaartorfiup Nunaa (magenta).

In terms of encounter rates, *i.e.*, the number of detections per unit transect length, the Sisimiut sub-area presented an average encounter rate of 1.84 caribou groups per km, while Sisimiut South and Angujaartorfiup Nunaa sub-areas registered mean encounter rates of

0.45 and 0.63 caribou groups per km, respectively (Figure 4.3). The remarkably constant encounter rate leads to reasonable precision within the data.

Regarding the effect of the observer, it appears that this covariate does not seem very helpful in explaining detectability as both histograms are similar (Figures 4.4, and Figure 5.3 in the Appendices).

As previously said, there would be other potential available covariates, like sun glare or snow covering, but there were too many missing observations and/or inconsistency when referring to the categories for these to be used in the analysis.

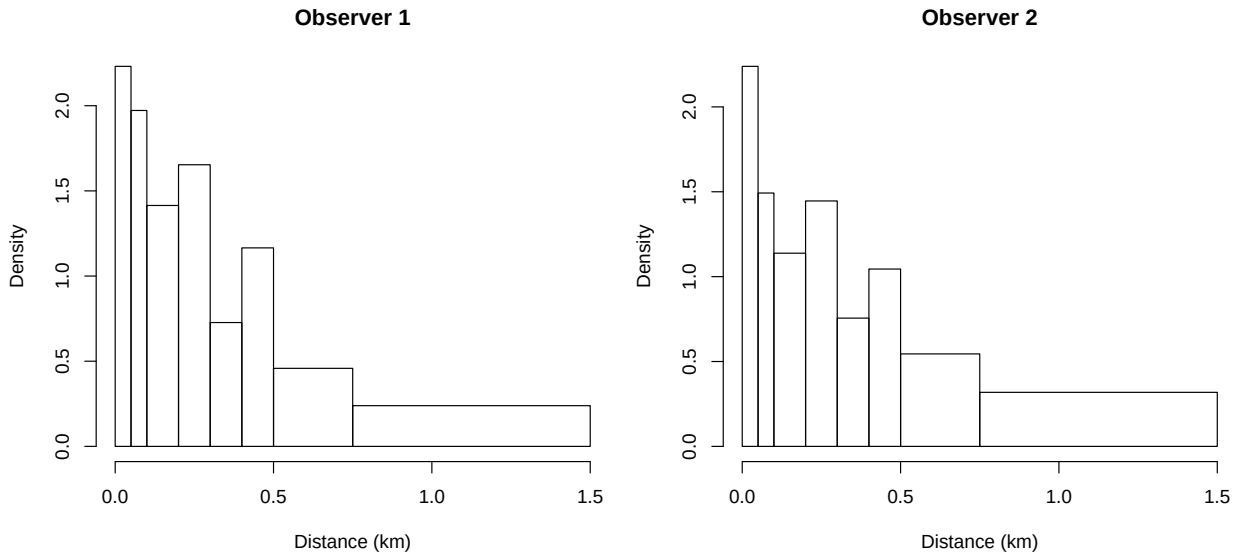

Figure 4.4: The detected distances' histogram for each observer (a covariate with two levels).

Overall, this exploratory analysis anticipates a good precision since the information is in accordance for each sub-area and with what was expected *a priori*, with Sisimiut being the sub-area expected to have more caribou when compared with the other two sub-areas.

4.3 Distance Sampling

Initially, an analysis of the observed distances must be made to evaluate if there is any major assumption violation or other data-related issues stated in the previous chapters before conducting the modelling step. The histogram of observed distances with no defined truncation distance is similar to typical Distance Sampling data, perhaps showing some over-dispersion, with not-equally-spaced bins (Figure 4.5).

After observing the histogram of binned distances, a strip half-width of $w = 0.75$ km was selected, reducing the sample size to 1640 caribou groups. Data truncation is a common procedure in this type of analysis due to the fact that otherwise extra adjustment terms may be needed to fit the long tail of the detection function. Also, by doing this, little information

is lost since data observations located more than 0.75 km from each side of the line make little contribution to the abundance estimate.

An alternative binning option (Option 2), with less bins, reduces the influence of potential measurement errors in the observed distances. This alternative binning option includes bin cutpoints of 0, 0.10, 0.30, 0.50 and 0.75 km (Figure 4.5).

With the original binning option, there seem to be less than expected observations on the 0.10-0.20 km and 0.30-0.40 km intervals, when compared to the 0.20-0.30 km and the 0.40-0.50 km bins. This might be an evidence of *heaping*. This phenomenon, as said before, occurs when observers tend to record some preferred values over others (Buckland et al., 2001). In this case, the heaping would have occurred for distances 0.25 km and 0.50 km, which are round distances that are easily picked in the absence of a rigorous distance measuring method.

Both binning options were considered in model fitting, despite the fact that only the second option minimizes the effect of measurement error induced by heaping. Below, only the analyses whose fitted models consider the second binning option are shown, since the former binning option is not suitable for grouping. Some details of the models fitted with the first binning option are show in the Appendices (Table 5.7, and Figure 5.8).

The advantage of choosing the second binning option is that it results in more reliable detection functions. Meanwhile, the small number of bins impacts on the Goodness-of-Fit tests after model fitting, since it results in a lower number of degrees of freedom.

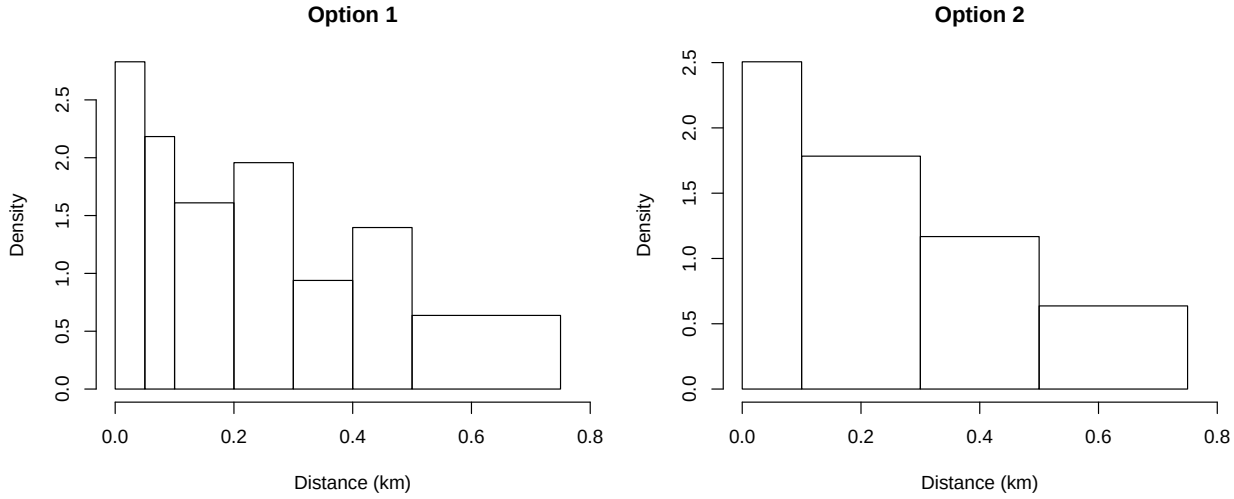

Figure 4.5: Histogram of the different binning options for the caribou distance data. Left: the original bins as collected on the survey. Right: an alternative binning to reduce the effect of heaping. The area of the rectangles is proportional to the number of points within each bin.

A scatter plot, with a GLM fitted between two variables, observed distance as explanatory variable, and group size as response variable, suggested a faint tendency for larger groups being associated with greater distances (Figure 4.6). Note that the maximum group size is no

longer 20 as this data set is truncated, considering a strip width of $w = 0.75$ km, therefore, the most distant observations, which corresponded to large group sizes, were excluded. This regression analysis suggests that distance is a statistically significant variable explaining group size (Tables 4.1, and Table 5.2 in the Appendices). Group size also seemed to be related with the spatial coordinates (Figure 5.4 in the Appendices).

Table 4.1: Summary coefficient characteristics of the GLM between observed distance and group size and considering a Poisson distribution.

	Estimate	Std. Error	z value	p-value
Intercept	0.747	0.029	26.20	0.00000
Distance	0.311	0.080	3.88	0.00011

Note:
AIC = 5613.7, Null Deviance = 1361.4
Residual Deviance = 1346.4

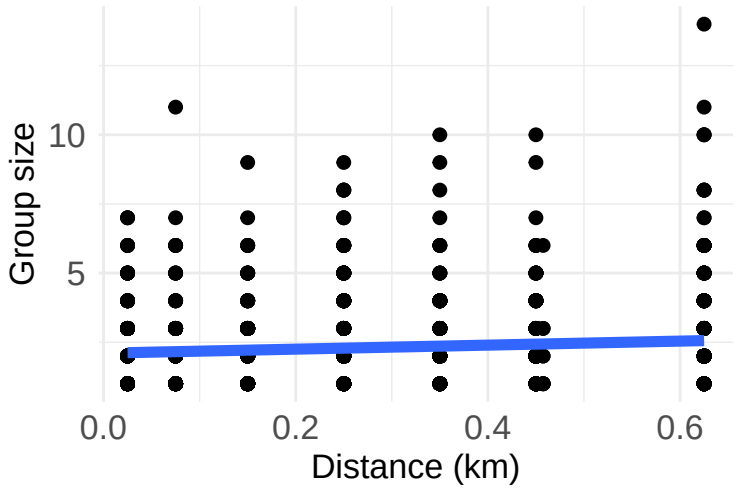

Figure 4.6: Relationship between group size and observed distances and respective regression fit using GLM.

Detection function models fitted with the first binning option did show poor fitting, including the best fit within this group, since these presented several adjustment terms, due to the heaping phenomena. Below, the detection functions are fitted to the data considering the second binning option.

For these models, every combination of key function and adjustment terms was tested. Only observer and group size were the additional covariates assessed, considering $w = 0.75$ km.

A summary of the information from each model fitted to the data is usually provided (Table 4.2). It allows a simple overview of a large number of models since it includes the respective key functions, adjustment terms, model formula, χ^2 Goodness-of-Fit test p-value, estimates of

the detection probability, respective standard error ($se(\widehat{P}_a)$), and ΔAIC comparison between each model and the model with the lowest AIC. The best model fitted to the data is bearing the lowest AIC value ($\Delta AIC = 0$). For this data, this model has the hazard rate function as a key function, no adjustment terms added and only group size as covariate (AIC = 4414.52). The hazard rate key function was selected maybe due to the fact of being the most flexible key.

Table 4.2: Model comparison across the three Conventional Distance Sampling models and models considering group size and observer as covariates.

Key function	Formula	χ^2 p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC
Hazard-rate	size	NA	0.541	0.025	0.000
Half-normal	size	0.000	0.603	0.013	7.658
Half-normal with cosine adjustment terms of order 2,3	1	NA	0.512	0.026	8.582
Uniform with cosine adjustment terms of order 1,2,3	NA	NA	0.513	0.025	8.582
Hazard-rate with cosine adjustment term of order 2	1	NA	0.519	0.025	8.647
Hazard-rate with simple polynomial adjustment term of order 2	1	NA	0.533	0.032	10.123
Hazard-rate	Obs	NA	0.544	0.025	10.672
Hazard-rate with Hermite polynomial adjustment term of order 4	1	NA	0.535	0.032	10.679
Uniform with simple polynomial adjustment terms of order 2,4,6	NA	NA	0.577	0.029	15.371
Half-normal	Obs	0.000	0.605	0.013	15.635
Half-normal	1	0.001	0.606	0.013	19.097
Half-normal	1	0.001	0.606	0.013	19.097
Uniform with Hermite polynomial adjustment term of order 4	NA	0.000	0.643	0.010	30.245

Note: Formula = 1 for Uniform key, NA otherwise (no covariates) Not enough d.f. for the GOF test, thus the 'NA' values.

The second best model includes the half-normal key with group size as a covariate (AIC = 4422.17, *i.e.*, $\Delta AIC = 7.66$). Therefore, there is strong evidence suggesting group size as a relevant covariate in detectability.

The best fitted detection function parameters' estimates shows a slight positive relationship between group size and detectability, superimposed with the observed distances' histogram (Table 4.3, and Figure 4.7). The estimated averaged probability of detection for the North region was $\widehat{P}_a = 0.541$ ($se = 0.025$). Remaining detection functions and summary table are in the Appendices (Figure 5.7, and Table 5.6). It is an averaged estimate since group size is included in the model. Consequently, each group size has its separate detection function, corresponding to different estimates for the probability of detection (Figure 4.8).

Table 4.3: Detection function parameters' estimates.

	Estimate	Std. Error
Intercept	-1.681	0.154
Group size	0.152	0.048

Note: Estimates on log scale.

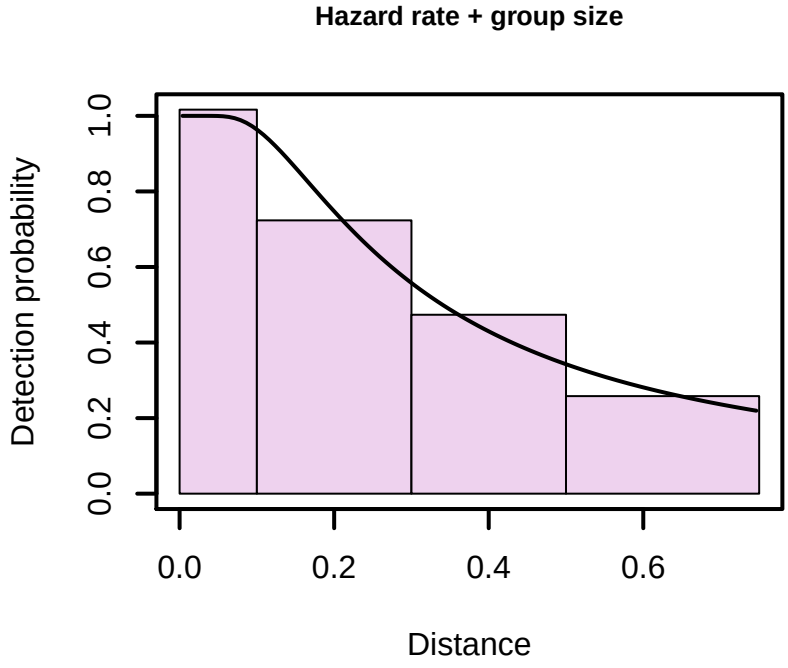

Figure 4.7: The detected distances with the estimated detection function overlaid, considering the binning option that reduces the effect of heaping.

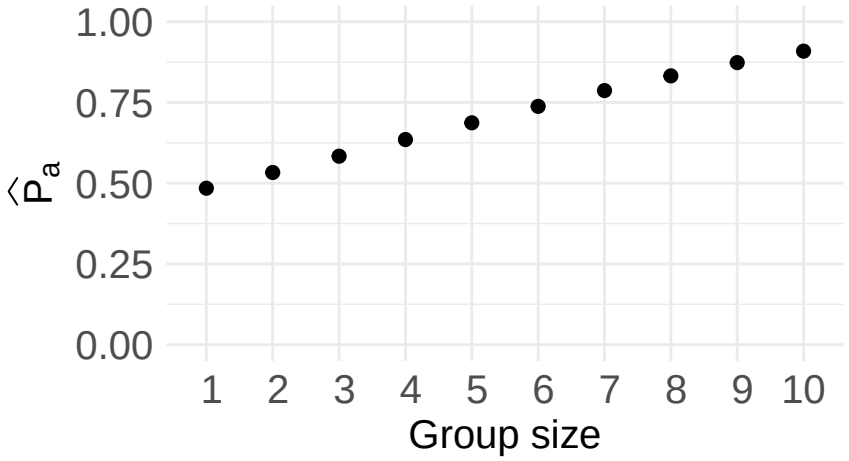

Figure 4.8: Estimated probabilities of detection for each observed group size obtained with the fitted model.

A group size of 2 caribou presents an estimated probability of detection of 0.533, while a group size of 10 has an estimate of 0.909. With increasing group size, the probability of detection also increases. This is consistent with the biologist's intuition as larger groups are easier to detect than smaller ones.

The estimates concerning the encounter rate suggest the Sisimiut sub-area to be the one with most caribou, since its estimate is larger compared to the other sub-areas (Table 4.4). Concerning the design-based estimates for caribou abundance and density, Sisimiut is also the sub-area presenting more caribou (Table 4.5 for abundance, and Figure 4.9 and Table 5.4 in the Appendices, for density).

Table 4.4: Encounter rate (ER) estimates per stratum for caribou groups considering three strata, five bins, and a detection function fitted with group size as a covariate.

Sub-area	ER	se(ER)	cv(ER)
Sisimiut	1.389	0.084	0.060
Sisimiut South	0.403	0.110	0.273
Angujaartorfiup Nunaa	0.559	0.085	0.152
Total	0.997	0.120	0.120

Note: se - standard error, cv - coefficient of variation.

The GOF test cannot be performed to the selected model since there are not enough degrees of freedom (from Equation (3.19), $u - q - 1 = 4 - 3 - 1 = 0$ degrees of freedom, and Table 5.5 from the Appendices presents the observed and expected values). Additionally, the Kolmogorov-Smirnov and Cramér-von Mises tests cannot be applied, since the distances are represented as a discrete variable.

The design-based density estimate for the whole survey region was 2.59 (95% CI: (2.23, 3.02)) caribou per km^2. The Appendices contain further details (Table 5.4).

Table 4.5: Abundance estimates per stratum for caribou considering three strata, five bins and a Hazard rate detection function with group size as a covariate.

Sub-area	Estimate	Std. Error	cv	lcl	ucl
Sisimiut	46724	3745	0.080	39392	55422
Sisimiut South	3931	1134	0.289	1820	8492
Angujaartorfiup Nunaa	9814	1502	0.153	6758	14252
Total	60469	4501	0.074	51932	70410

Note: lcl, ucl - lower and upper 95% confidence limits, respectively.

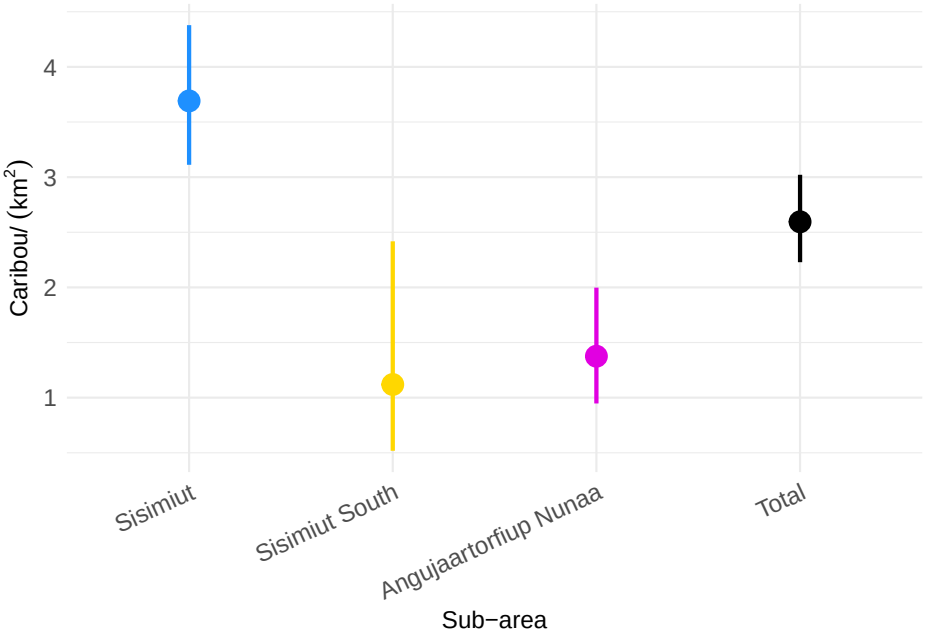

Figure 4.9: Caribou density estimates with corresponding confidence intervals for the sub-areas and entire North region.

It is estimated that, in the North region, there would be approximately 60 469 (95% CI: 51 932, 70 410) caribou, with a coefficient of variation (cv) of 7.4% which is a really low value when compared with previous studies' conclusions since it indicates relatively accurate estimates. Comparing with the previous results in Chapter 2 (Figure 2.3), it indicates a 38.5% decline in population size since 2010.

4.4 GAM and DSM

After splitting the transects into segments, their centroids were determined and intersected with the shapefiles concerning the spatial variables. These variables were GPS coordinates, vegetation, elevation, slope, and the aspect (geographical compass direction, related with sun

exposure and temperature of a particular location). The vegetation covariate was excluded from the analysis since the current information available lacks the necessary resolution. Furthermore, a prediction data set was generated for the whole study region by converting the region into small 1.5km side squared cells (*i.e.*, 2.25km^2, note $w = 0.75$km).

Once the distance model is fitted, $\widehat{P}_a$ can be determined for each group size and thus n_i, the number of caribou detected in group i, can be corrected as $\hat{n}_i = \frac{n_i}{\widehat{P}_a}$. These predicted values for counts were then modelled using a GAM fitted to the spatial covariates along with longitude and latitude considered jointly. Within the model, the previously estimated probability of detection was defined as the offset term along with the cell area of 2.25km^2. The distributions considered in model fitting were Tweedie and Negative Binomial. These provide a flexible alternative to the quasi-Poisson distribution which does not capture the response overdispersion.

Table 4.6: GAM model summary table relative to smooth terms for the covariates and considering a Tweedie distribution.

	EDF	Chi-square	p-value
s(lon, lat)	22.931	7.903	0.0000
s(Aspect)	5.503	5.580	0.0000
s(Elevation)	5.219	12.793	0.0000
s(Slope)	4.024	3.517	0.0036

Note:
EDF - Effective Degrees of Freedom.
$R^2_{adj} = 0.318$, Deviance explained = 38.7%

Table 4.7: GAM model summary table relative to smooth terms for the covariates and considering a Negative Binomial distribution.

	EDF	Chi-square	p-value
s(lon, lat)	22.468	153.876	0.0000
s(Aspect)	6.769	37.514	0.0000
s(Elevation)	4.665	74.290	0.0000
s(Slope)	4.383	18.731	0.0032

Note:
EDF - Effective Degrees of Freedom.
$R^2_{adj} = 0.278$, Deviance explained = 33.7%

Despite the similarity between the results, the model considering the Tweedie distribution was chosen since it presented a lower AIC value ($AIC_{Tweedie} = 5052$ and $AIC_{NegBinom} = 5453.9$). For the bivariate smooth associated with the pair longitude and latitude a larger EDF resulted, comparing to the environmental covariates, since more basis functions are

required to fit a surface than a line. Each of these environmental covariates does not appear to be linearly related with the response, since the respective EDF is larger than 1 (Tables 4.6 and 4.7). Apropos the smooth functions, the relationship between the response variable and each explanatory variable appears to be non-linear, given the other associations. South-facing (90º-270º) surfaces appear to be preferred over North-facing (270º-90º) ones. Regarding the slope, very steep locations seem to be highly avoided by caribou while those with less slope tend to be preferred. Concerning the elevation, caribou abundance seemed to exhibit a negative relationship with this covariate (Figure 5.5 in the Appendices). Furthermore, a faint tendency to prefer river areas can be noticed, but this may happen since these are low elevation areas. Therefore, lower altitudes are preferred by the study species, in the early March period for the survey, supporting previous results (Table 4.4) (Cuyler et al., 2017).

The number of caribou within each cell was then predicted using the GAM model ($\hat{n}_i$) and spatially represented. Overall, in accordance with the results provided by the design-based approach related with the sub-areas, the caribou seem to be more abundant in Sisimiut and less abundant in Angujaartorfiup Nunaa (Figure 4.10). The result appears related to elevation and those two regions differ in their relative proportion of low and high elevations, with Angujaartorfiup Nunaa having predominantly high elevations while low elevations are typically used by caribou in March (Cuyler et al., 2017).

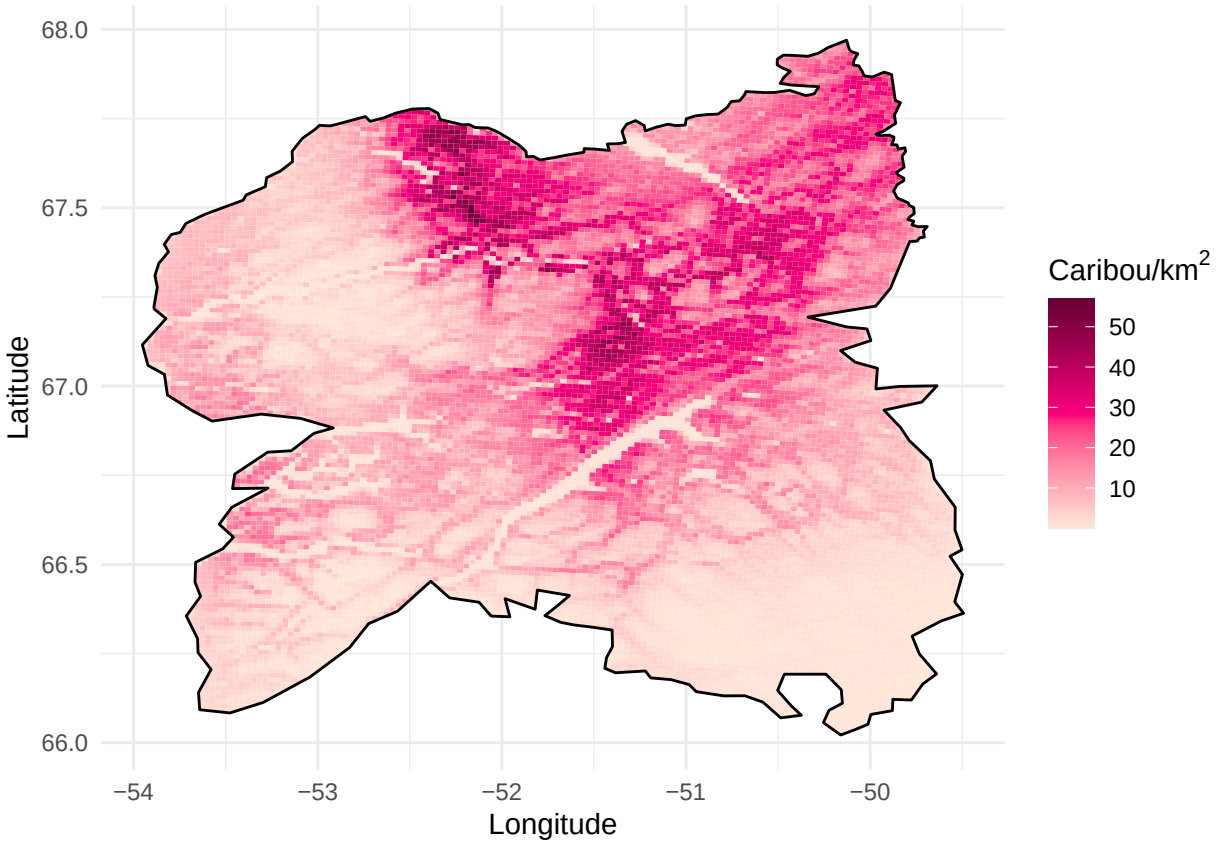

Figure 4.10: Predicted caribou abundance distribution across the entire North region.

The heat map presented for the prediction of caribou abundance in the North region is in accordance with previous studies (Cuyler et al., 2011) and predictions, with a higher density stratum being found in the northern part of the region (Sisimiut sub-area), while the Angujaartorgfiup Nunaa sub-area is characterized by a lower density.

As previously said, the MCDS model produced an estimate of 60 469 (95% CI: 51 932, 70 410) caribou for the North region. In contrast, the model-based approach produced an estimate of 73 895 (95% CI: 65 983, 82 757) caribou. While these differ by about 22%, we note the 95% CI overlap. This difference between both estimates may have resulted since the design-based methodology is based on the selected transects and these may not adequately represent every feature within the study region. Some features may have been over-represented, while others under-represented. On the other hand, the GAM modelling considers the environmental covariates in the whole region, allowing the spatial representation of the estimates obtained. This results in a better understanding and visualization of patterns in abundance. Additionally, a variability map was produced with the CV for each estimate in the survey region (Figure 5.6 in the Appendices). Overall, the CV estimate was 5.78%, pointing towards great precision within the analysis. The design-based approach had $\widehat{CV}$ = 4.55%, while the estimate for the GAM model was lower, $\widehat{CV}$ = 3.57%, suggesting an improvement within the analysis. Comparing with other studies (Buckland et al., 2015; Miller et al., 2016), the presented variability estimates point to great precision associated with the data and the estimates, which may also be due to the large sample size.

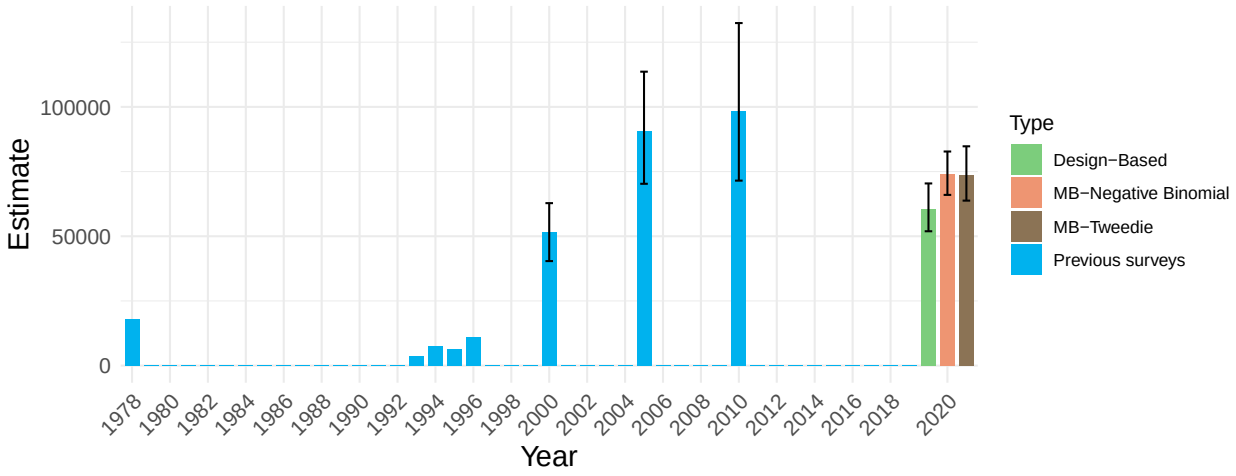

Figure 4.11: Abundance estimates for caribou in the North region obtained in previous surveys (blue), design-based estimates (green) and model based estimates for Negative Binomial (brown), and Tweedie models (orange). Black lines represent the confidence intervals (90% for previous surveys and 95% for the remaining).

Finally, comparing with previous results, a decrease in the overall abundance is suggested, even though the 95% CI overlap (Figure 4.11). These abundance estimates may be slightly biased since the observers likely have overestimated distances during the survey (C. Cuyler, pers. comm.). The bias is expected to push the population estimate downward. Considering

the estimates from each approach, there may be a slow decline in the caribou abundance within the North region.

Chapter 5

Conclusions

Since the caribou are an important resource in West Greenland, up-to-date abundance estimates are needed for management decisions aimed at controlling animal density. This project's results suggest a recent declining trend in the Kangerlussuaq-Sisimiut caribou population. Although new data and survey improvements have been made, further investigation about this subspecies is required.

The main goals of this project were to estimate the caribou abundance and spatial distribution in the North region of West Greenland. The MCDS abundance estimate was higher than the prediction using the GAM model, but the 95% CI overlapped.

In general, the differences between abundance estimates may be explained by the different techniques used within the analysis. With MCDS being a design-based approach, whose performance depends on the selected line transects. If these are preferentially located in a high density area, density will be overestimated and vice-versa, since the design was biased. Furthermore, the possible distance overestimation can lead to biased density estimates. On the other hand, GAM is a model-based approach with additional covariates related to density present for the whole study region. Unlike the design-based approach, the risk of certain features being under or over-represented within the data (*e.g.*, lakes) is lower. The model-based approach, through the covariates, considers the whole region of interest and not only the surveyed area, providing a more robust abundance estimate subject to the model being a reasonable approximation of reality.

Design-based and model-based point estimates point to a decrease in the caribou abundance since the 2000-2010 period, but this difference is not statistically signifficant since the CI overlap. This was the successful outcome of management directed at avoiding caribou overabundance. Regarding the spatial distribution, it also agreed with the specialists' predictions (C. Cuyler, pers. comm), where higher abundance could be found in the upper half of the region, specifically, the Sisimiut sub-area.

The estimates presented might also be affected by potential sources of bias. Some covariates could not be included in the analysis since these had an excessive number of missing observations or had inconsistent categorization. More covariates could be added, namely vegetation. The selected transects do not pose a problem since these were maximized as per

the provided budget and precisely allocated to avoid gradients (climate, vegetation, animal density and geographic/topographic), providing great coverage.

Finally, as a suggestion for future surveys, if possible, stricter data collection procedure in terms of categorization of variables of interest such as vegetation/snow coverage, snow depth and glare. This would generate more detailed information through the years and other patterns may arise for this important resource subspecies.

Acquired skills

At a personal level, this project has taught me what it is to be a statistician. It allowed me to expand my knowledge, to study new methods that were never taught through the Master's course in Biostatistics and to be able to apply them into (and to solve) real applied research questions. Many skills were acquired while working on this project and they are worth mentioning here. Throughout this project I could:

1. learn how to work in a real scientific applied research topic;
2. develop my programming skills;
3. sharpen problem-solving skills and critical thinking;
4. boost self-confidence both in programming and statistical reasoning;
5. be included in a project that allowed me to implement different methodologies both learned during the Master's course and the project (exploratory analysis, Distance Sampling, model selection, spatial data, GAM, DSM);
6. learn about useful packages specifically **bookdown** and **knitr**, allowing me to strengthen my knowledge with R Markdown and LaTeX for both statistical analyses and mathematical writing.

Altogether, these skills acquired throughout the project will certainly help with the future that I intend to pursue, expand my knowledge and grow as a biostatistician.

Appendices

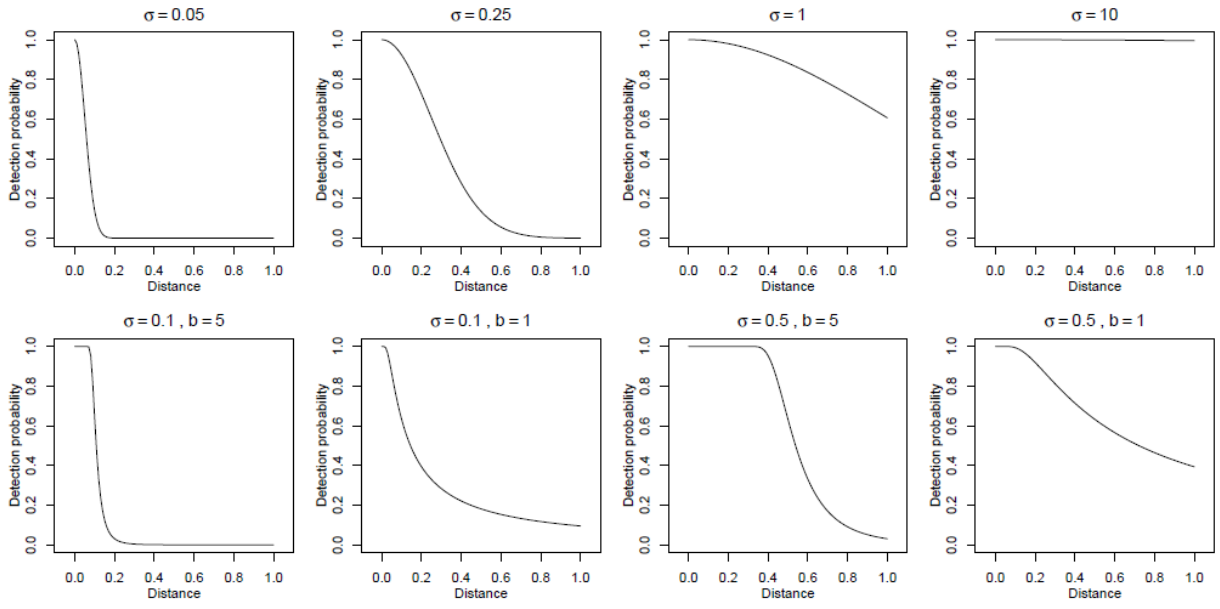

Figure 5.1: Half-normal (top row) and hazard-rate (bottom row) detection functions without adjustments, varying scale (σ) and, only for hazard-rate, shape (b) parameters. Values tested are presented above the plots. On the top row from left to right, the study species becomes more detectable (higher probability of detection at larger distances). The bottom row shows the hazard-rate model's more pronounced shoulder. Adapted from Buckland et al. (2001).

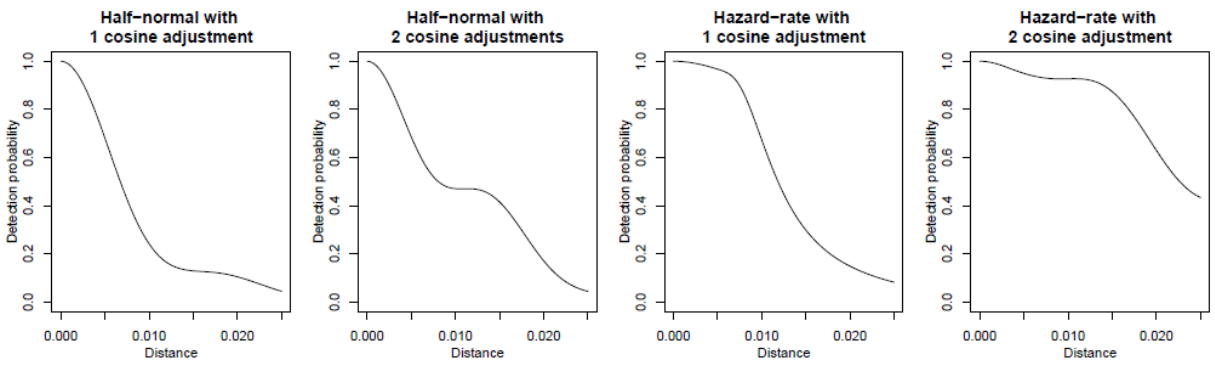

Figure 5.2: Possible shapes for the detection function when cosine adjustments are included for half-normal and hazard-rate models. Adapted from Buckland et al. (2001).

Table 5.1: Late winter caribou parameters from aerial and ground surveys of the North region. Adapted from Cuyler et al. (2011).

	1998	2000	2005
Mean group size (SD)	2.3	2.7	4.6 (3.4)
Maximum group size	10	17	17
Density		1.2 to 2.8	2 to 6

Note: The provided densities are in caribou per km^2.

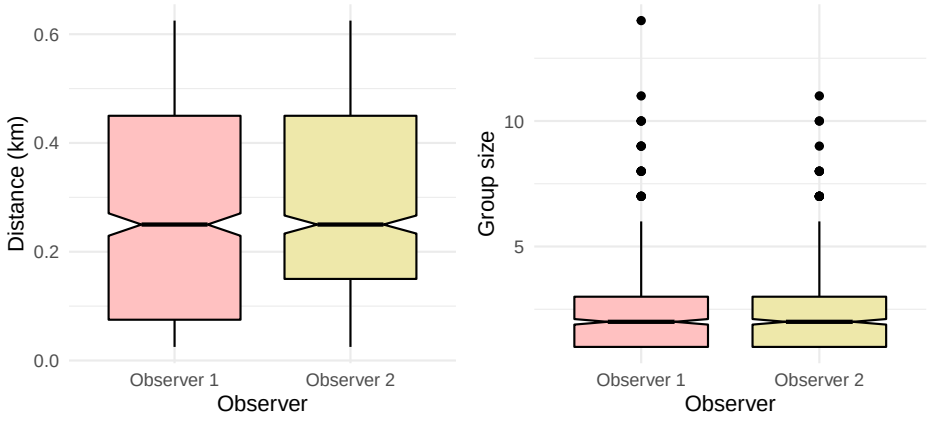

Figure 5.3: Boxplots of the detection distances for both observers present during the aerial survey. The boxes' upper and lower limits correspond to the first and third quartiles, respectively, and the centre line being the second quartile (median). The notch displays the 95% confidence interval around the median (M), based on $M \pm 1.58 \cdot IQR/\sqrt{n}$.

Table 5.2: ANOVA table of the Generalized Linear Model fitting.

	df	Deviance	Residual df	Residual Deviance
Null			1639	1361.36
Distance	1	14.91	1638	1346.45

Note: df - degrees of freedom.

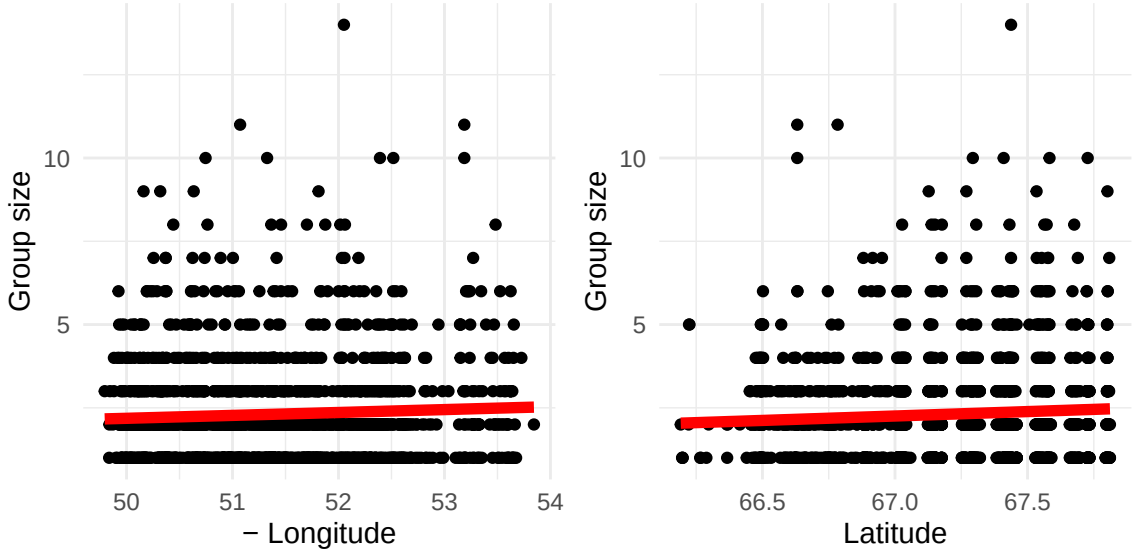

Figure 5.4: Scatter plot and respective univariate GLM fit between group size and longitude and latitude, respectively.

Table 5.3: Survey and encounter rate (ER) statistics.

Sub-area	Area	Covered Area	Effort	n	ER	cv(ER)	Mean size
Sisimiut	12658	1374.705	916.47	2974	3.245	0.069	2.336
Sisimiut South	3512	387.090	258.06	260	1.008	0.291	2.500
Angujaartorfiup Nunaa	7133	705.690	470.46	551	1.171	0.144	2.095
Total	23303	2467.485	1644.99	3785	2.301	0.128	2.308

Note: n - sample size, cv - coefficient of variation.

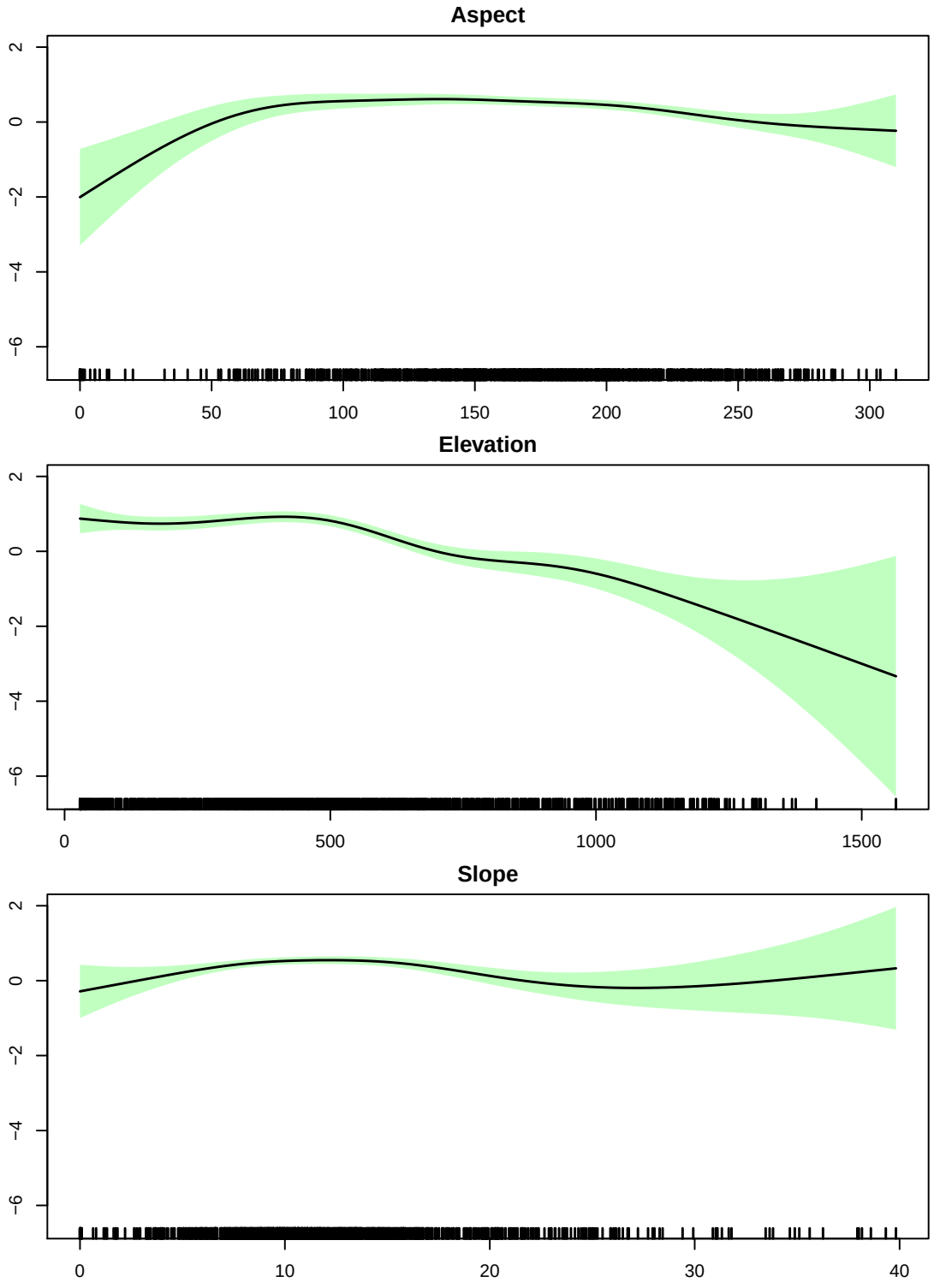

Figure 5.5: Fitted DSM plots. The plots correspond to the fitted smooths of the covariates included in the model, aspect, elevation and slope, respectively. The green shaded area is the standard error of the estimates, while the respective observations are represented on the horizontal axis.

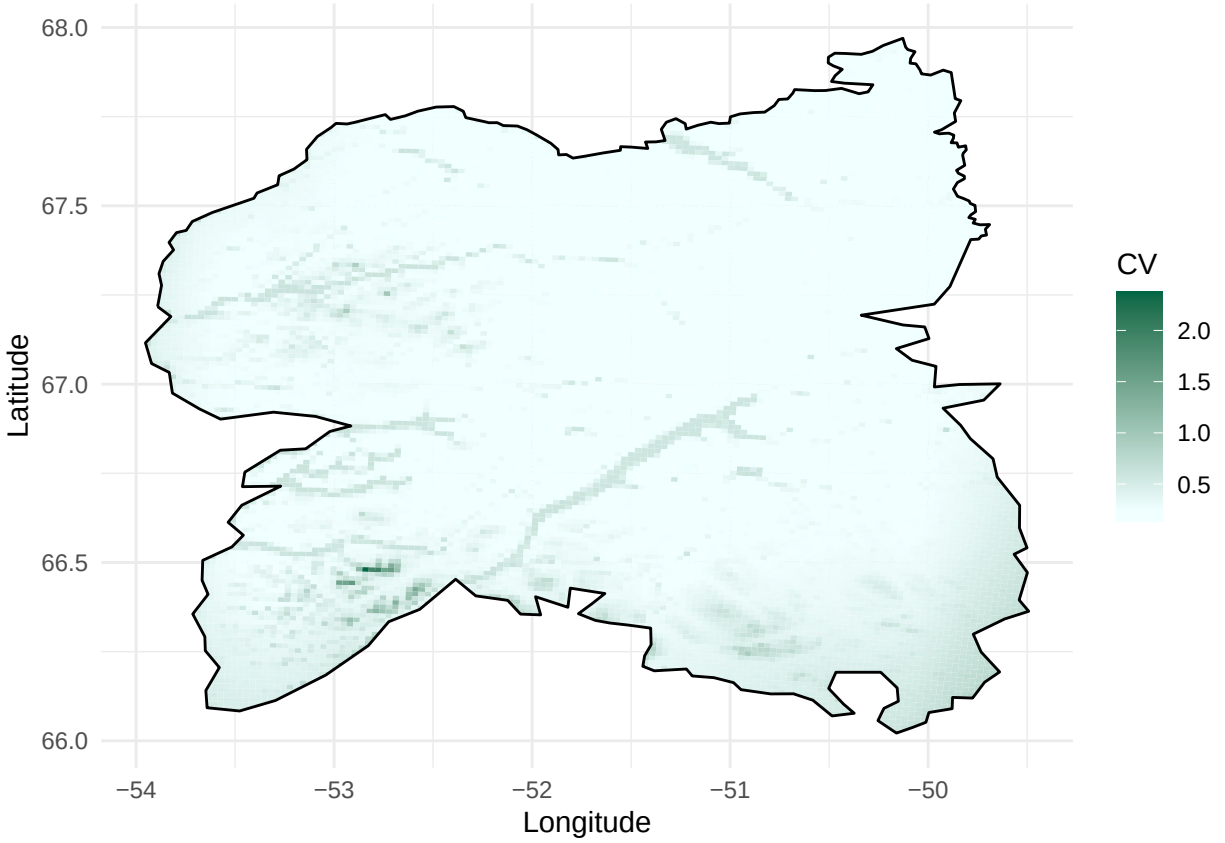

Figure 5.6: Fitted DSM variability plot. The darker the region, the higher the CV, that is, the larger the uncertainty associated to the respective density estimate.

Table 5.4: Density estimates per stratum for caribou considering three strata, five bins and a Hazard rate detection function with group size as a covariate.

Sub-area	Estimate	Std. Error	cv	lcl	ucl
Sisimiut	3.69	0.296	0.080	3.11	4.38
Sisimiut South	1.12	0.323	0.289	0.52	2.42
Angujaartorfiup Nunaa	1.38	0.211	0.153	0.95	2.00
Total	2.59	0.193	0.074	2.23	3.02

Note: lcl, ucl - lower and upper 95% confidence limits.

Table 5.5: Chi-square Goodness-of-Fit test observed and expected values for the best Distance Sampling model considering the second binning option.

	[0, 0.1]	(0.1, 0.3]	(0.3, 0.5]	(0.5, 0.75]
Observed	411.00	585.0	383.00	261.00
Expected	401.95	609.5	352.94	275.61

Table 5.6: Model comparison across the three Conventional Distance Sampling models and models considering group size and observer as covariates. This table was intentionally repeated to work as a guidance for the detection functions' plots for the remaining models considering the second binning option.

Model	Key function	Formula	χ^2 p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC
dsCARs22bins2	Hazard-rate	size	NA	0.541	0.025	0.000
dsCARs12bins2	Half-normal	size	0.000	0.603	0.013	7.658
dsCar12bins2.1	Half-normal with cosine adjustment terms of order 2,3	1	NA	0.512	0.026	8.582
dsCar32bins2.1	Uniform with cosine adjustment terms of order 1,2,3	NA	NA	0.513	0.025	8.582
dsCar22bins2.1	Hazard-rate with cosine adjustment term of order 2	1	NA	0.519	0.025	8.647
dsCar22bins2.3	Hazard-rate with simple polynomial adjustment term of order 2	1	NA	0.533	0.032	10.123
dsCARObs22bins2	Hazard-rate	Obs	NA	0.544	0.025	10.672
dsCar22bins2.2	Hazard-rate with Hermite polynomial adjustment term of order 4	1	NA	0.535	0.032	10.679
dsCar32bins2.3	Uniform with simple polynomial adjustment terms of order 2,4,6	NA	NA	0.577	0.029	15.371
dsCARObs12bins2	Half-normal	Obs	0.000	0.605	0.013	15.635
dsCar12bins2.2	Half-normal	1	0.001	0.606	0.013	19.097
dsCar12bins2.3	Half-normal	1	0.001	0.606	0.013	19.097
dsCar32bins2.2	Uniform with Hermite polynomial adjustment term of order 4	NA	0.000	0.643	0.010	30.245

Note: Formula = 1 for Uniform key, NA otherwise (no covariates) Not enough d.f. for the GOF test, thus the 'NA' values.

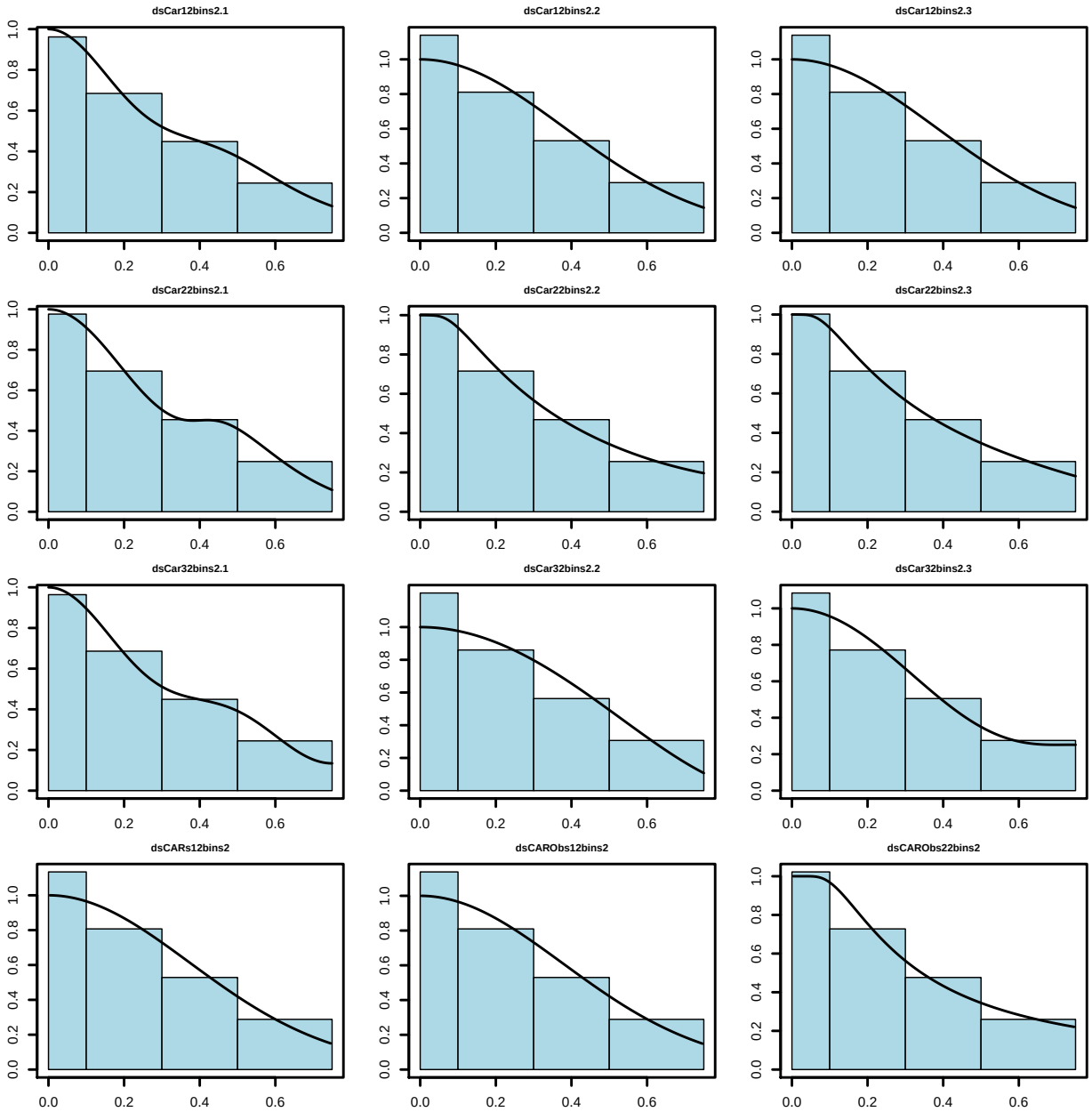

Figure 5.7: The detected distances' histograms superimposed with the estimated detection functions for the second binning option, excluding the best fit.

Table 5.7: Model comparison across the three Conventional Distance Sampling models and models considering group size and observer as covariates, with the first binning option (bin cutpoints at 0, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50 and 0.75 km).

Model	Key function	Formula	χ^2 p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC
dsCar22.1	Hazard-rate with cosine adjustment term of order 2	1	0	0.547	0.020	0.000
dsCARs22	Hazard-rate	size	0	0.551	0.025	3.380
dsCar12.1	Half-normal with cosine adjustment terms of order 2,3	1	0	0.535	0.023	4.208
dsCARs12	Half-normal	size	0	0.620	0.014	4.689
dsCARObs12	Half-normal	Obs	0	0.622	0.014	12.285
dsCar32.3	Uniform with simple polynomial adjustment terms of order 2,4,6,8,10	NA	0	0.588	0.036	12.764
dsCar22.2	Hazard-rate with Hermite polynomial adjustment term of order 4	1	0	0.563	0.030	13.844
dsCar32.1	Uniform with cosine adjustment term of order 1	NA	0	0.617	0.012	14.462
dsCar12.2	Half-normal	1	0	0.622	0.014	14.769
dsCar12.3	Half-normal	1	0	0.622	0.014	14.769
dsCARObs22	Hazard-rate	Obs	0	0.560	0.025	16.889
dsCar32.2	Uniform with Hermite polynomial adjustment term of order 4	NA	0	0.651	0.011	20.791

Note: Formula = 1 for Uniform key, NA otherwise (no covariates) Not enough d.f. for the GOF test, thus the 'NA' values.

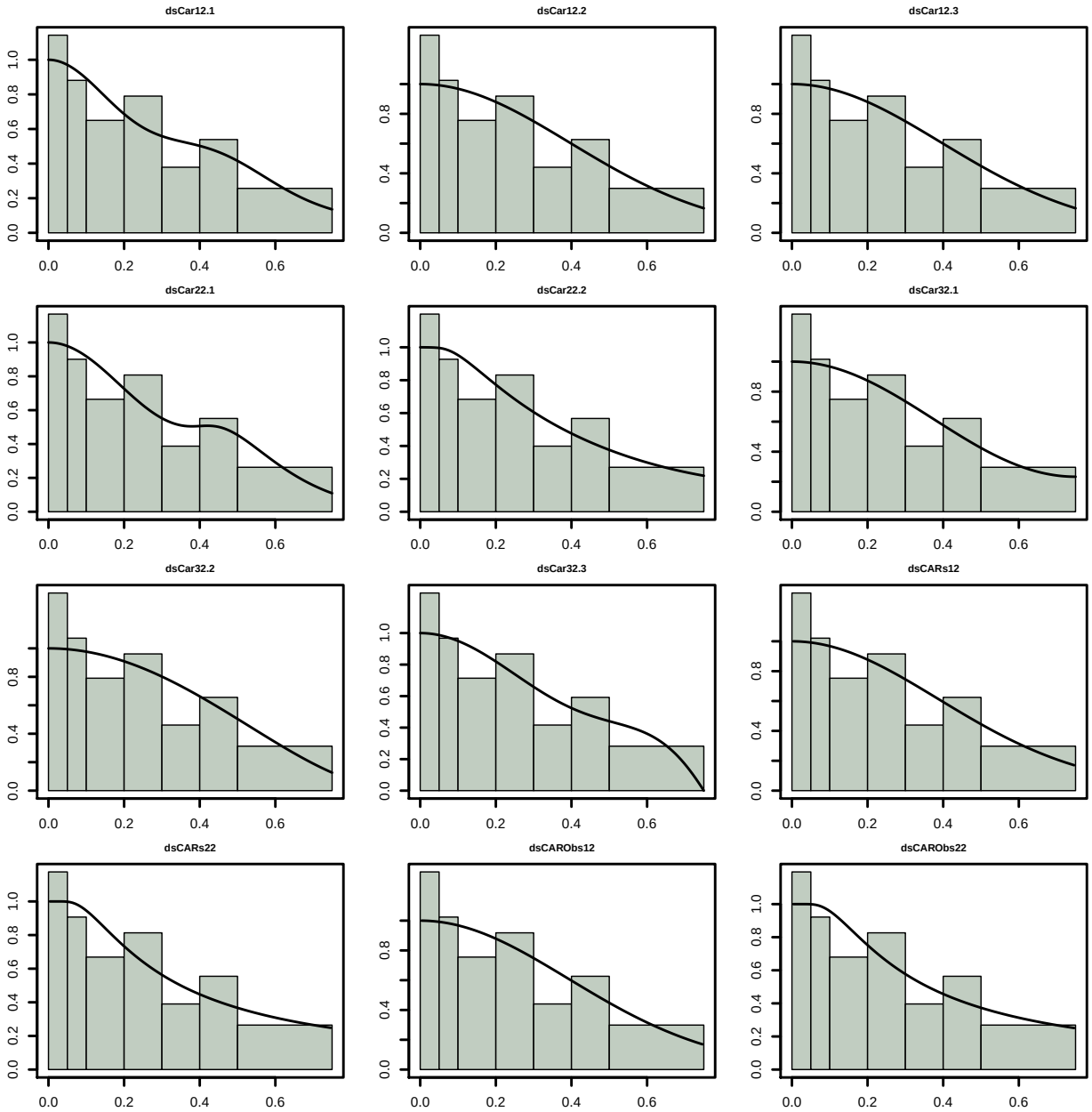

Figure 5.8: The detected distances' histograms superimposed with the estimated detection functions for the first binning option.

Bibliography

Anderson, D. R., Burnham, K. P., Lubow, B. C., Thomas, L., Corn, P. S., Medica, P. A., and Marlow, R. W. (2001). Field trials of line transect methods applied to estimation of desert tortoise abundance. *Journal of Wildlife Management*, 65(3):583–597.

Brewer, M. J., Butler, A., and Cooksley, S. L. (2016). The relative perperformance AIC, AICc and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6):679–692.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations*. Springer.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001). *Introduction to Distance Sampling Estimating abundance of biological populations*. Oxford University Press.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2004). *Advanced Distance Sampling: Estimating abundance of biological populations*. Oxford University Press.

Buckland, S. T., Rexstad, E. A., Marques, T. A., and Oedekoven, C. S. (2015). *Distance Sampling: Methods and Applications*. Springer.

Couturier, S., Dale, A., Wood, B., and Snook, J. (2018). Results of a Spring 2017 aerial survey of the Torngat Mountains Caribou Herd. Technical report, Torngat Wildlife, Plants and Fisheries Secretariat.

Cuyler, C. (2007). West Greenland caribou explosion: What happened? What about the future? *Rangifer*, 27(4):219–226.

Cuyler, C., Nagy, J., and Zinglersen, K. (2017). Seasonal movementanda ctivity of Akia-Maniitsoq caribou cows in west greenland as determined by satellite. Technical report 99, Greenland Institute of Natural Resources, West Greenland.

Cuyler, C., Nymand, J., Jensen, A., and Molgaard, H. (2016). 2012 status of two West Greenland caribou populations 1) Ameralik, 2) Qeqertarsuatsiaat. Technical Report 98, Greenland Institute of Natural Resources, West Greenland.

Cuyler, C., Rosing, M., Egede, J., Heinrich, R., and Molgaard, H. (2005). Status of two West Greenland caribou populations 2005; 1) Akia-Maniitsoq, 2) Kangerlussuaq-Sisimiut. Technical Report 61, Greenland Institute of Natural Resources, West Greenland.

Cuyler, C., Rosing, M., Heinrich, R., Egede, J., and Mathaeussen, L. (2007). Status of two West Greenland caribou populations 2006, 1) Ameralik, 2) Qeqertarsuatsiaat. Technical Report 67, Greenland Institute of Natural Resources, West Greenland.

Cuyler, C., Rosing, M., Linnell, J., Lund, P., Jordhoy, P., Loison, A., and Landa, A. (2003). Status of three West Greenland caribou populations; 1) Akia-Maniitsoq, 2) Ameralik and 3) Qeqertarsuatsiaat. Technical Report 46, Greenland Institute of Natural Resources, West Greenland.

Cuyler, C., Rosing, M., Linnell, J. D. C., Loison, A., Ingerlsev, T., and Landa, A. (2002). Status of the Kangerlussuaq-Sisimiut caribou population (*Rangifer tarandus groenlandicus*) in 2000. Technical Report 42, Greenland Institute for Research of Natural Resources, West Greenland.

Cuyler, C., Rosing, M., Molgaard, H., Heinrich, R., and Raundrup, K. (2011). Status of two West Greenland caribou populations 2010; 1) Kangerlussuaq-Sisimiut, 2) Akia-Maniitsoq. Technical Report 78, Greenland Institute of Natural Resources.

Cuyler, C., Rosing, M., Molgaard, H., Heirinch, R., Egede, J., and Mathaeussen, L. (2009). Incidental observations of muskox, fox, hare, ptarmigan and eagle during caribou surveys in West Greenland. Technical Report 75, Greenland Institute of Natural Resources, West Greenland.

Gibbons, J. D. and Chakraborti, S. (2011). *Nonparametric Statistical Inferencing*. Chapman & Hall.

Katsanevakis, S. (2007). Density surface modelling with line transect sampling as a tool for abundance estimation of marine benthic species: the *Pinna nobilis* example in a marine lake. *Marine Biology*, 152:77–85.

Linnell, J. D. C., Cuyler, C., Loison, A., Lund, P. M., Motzfeldt, K. G., Ingerslev, T., and Landa, A. (2000). The scientific basic for managing the sustainable harvest of caribou and muskoxen in Greenland for the 21st century: an evaluation and agenda. Technical Report 34, Greenland Institute of Natural Resources, Pinngortitaleriffik.

Marques, T. A. (2009). Distance Sampling: estimating animal density. *Significance*, 6(3):136–137.

Marques, T. A. (2018a). Estimating caribou abundance for GINR s 2018 West Greenland caribou survey. Techincal Report 3, Centre for Research into Ecological and Environmental Modelling. Report produced for GINR under a research contract between CREEM and GINR.

Marques, T. A. (2018b). Estimating muskox abundance in the Angujaartorfiup Nunaa sub-area using data collected during GINR's 2018 West Greenland caribou survey. Technical Report 2, Centre for Research into Ecological and Environmental Modelling. Report produced for GINR under a research contract between CREEM and GINR.

Marques, T. A., Buckland, S. T., Borchers, D. L., Rexstad, E., and Thomas, L. (2011). Distance Sampling. *International Encyclopedia of Statistical Science*, 1:398–400.

Marques, T. A., Thomas, L., Fancy, S. G., and Buckland, S. T. (2007). Improving estimates of bird density using multiple covariate distance sampling. *The Auk*, 124(4):1229–1243.

Meldgaard, M. (1986). The Greenland Caribou - Zoogeography, Taxonomy and Population Dynamics. *Bioscience*, 20:14.

Miller, D. L., Burt, M. L., Rexstad, E. A., and Thomas, L. (2013). Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution*, 4(11):1001–1010.

Miller, D. L., Rexstad, E., Thomas, L., Marshall, L., and Laake, J. L. (2016). Distance Sampling in R. *Journal of Statistical Software*, 89(1):1–28.

Poole, K. G., Cuyler, C., and Nymand, J. (2013). Evaluation of caribou *Rangifer tarandus groenlandicus* survey methodology in West Greenland. *Wildlife Biology*, 19:225–239.

Tamstorf, M. P., Aastrup, P., and Cuyler, C. (2005). Modelling critical caribou summer ranges in West Greenland. *Polar Biology*, 28:714–724.

Thomas, L., Buckland, S. T., Burnham, K. P., Anderson, D. R., Laake, J. L., L., B. D., and Stindberg, S. (2002). Distance Sampling. *Encyclopedia of Environmetrics*, 1:544–552.

Thomas, L., Buckland, S. T., Rexstad, E. A., Laake, J. L., Stindberg, S., Hedley, S. L., Bishop, J. R. B., Marques, T. A., and Burnham, K. P. (2010). Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology*, 47(1):5–14.

Witting, L. and Cuyler, C. (2008). Harvest impacts on caribou population dynamics in South West Greenland. *Rangifer*, 31(2):135–145.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.19.